CHAPTER 1: SUMMARY AND CONCLUSIONS

The primary goals of this dissertation were to improve damping modeling for sounding objects and to ensure sensory cohesion in multimodal interaction. For damping modeling, I presented multiple methods for estimation of material damping properties from recorded impact sounds. These methods use expressive damping models to capture complex behavior and reduce the effects of external factors. For multimodal interaction, I presented methods for automatically modeling real-world objects and interacting with virtual objects. These methods demonstrate a benefit when using multiple modalities of interaction.

1.1 Summary of Results

In this dissertation, I presented four primary contributions. I presented an extension of Generalized Proportional Damping (GPD) to modal sound synthesis, along with a method for automatically estimating material damping parameters for any GPD-derived damping model (Footnote 1). I presented a method for estimating material damping parameters in the presence of confounding external factors by using a probabilistic damping model (Footnote 2). I presented methods for multimodal interaction with textured surfaces while ensuring sensory cohesion (Footnote 3). I presented a method for producing estimates of object shape and material by fusing audio and visual input (Footnote 4).

While previous methods for modal sound synthesis only used the common Rayleigh damping model, my coauthors and I extended the traditional synthesis framework to support multiple damping models. We further used Generalized Proportional Damping to propose additional physically-based damping models. We presented a novel method for estimating the damping parameters of any of these damping models from a recorded impact sound. These damping models and the estimation method were evaluated in a perceptual study. No damping model was found to be consistently superior to the others, demonstrating that Rayleigh damping alone is not capable of expressing the full variability in damping behavior.

My coauthors and I further extended damping modeling to a probabilistic framework. This framework can extend any traditional damping model, while probabilistically modeling external effects such as background
noise, support damping, and error in feature extraction. For comparison purposes, we presented a study evaluating human performance in manual hand-tuning of damping parameters, finding that manual estimation requires significant human effort. In a perceptual evaluation of our method, we found that sounds synthesized using our parameters produce a pattern of errors similar to that observed in sounds synthesized using the hand-tuned parameters, indicating high perceptual similarity. Compared to parameters from Ren et. al (Ren et al., 2013), our synthesized sounds are perceptually more similar to recorded sounds on three out of four quality metrics. Overall, our method significantly reduces required human effort while producing sounds that are perceptually similar to recorded or hand-tuned sounds.

To address issues of sensory conflict in interactive multimodal environments, my coauthors and I evaluated multimodal surface interaction. We presented a method for multimodal interaction using texture maps as a unified representation of fine surface detail, performing evaluation through two user studies. Our texture identification user study found that users perceived texture identification to be easiest to perform when all modalities of interaction were provided. Our study comparing normal and relief maps found that the perceived realism of interaction with relief-mapped surfaces was higher than that of normal mapped surfaces when considering all modalities of interaction. When each modality of interaction was considered independently, normal maps alone were sufficient. These results suggest that multimodal interaction using unified representations of detail can improve sensory cohesion.

To aid in virtualization of objects and scenes for multimodal interaction, my coauthors and I presented a method for object shape and material identification from combined audiovisual input. Our Impact Sound Neural Networks (ISNN-A for audio-only input and ISNN-AV for combined audio-visual input) can provide accurate identification and classification of object shapes and geometry, using sound to improve accuracy in the cases of transparent or occluded objects. Accuracy is further improved by the use of spectrogram inputs and synthetic impact sounds to augment training. For audio-only object identification tasks, our ISNN-A network outperforms alternate models such as SoundNet (Aytar et al., 2016), achieving up to 99.52 % accuracy on synthetic datasets and up to 92.37 % accuracy on recorded real-world datasets. On the ModelNet (Kanezaki et al., 2016) datasets, our multimodal ISNN-AV network provide an improvement over the visual-only VoxNet (Maturana and Scherer, 2015), achieving up to 91.8 % accuracy.
1.2 Limitations

While my proposed methods improve the expressiveness of object damping modeling and improve multimodal interaction with virtualized objects, some limitations remain. While the limitations of each method is discussed in its respective chapter, these limitations relate primarily to the general methodology and assumptions made across my work.

First, the proposed methods use linear models for object vibrations. While linear models are critical for achieving real-time runtime synthesis, impact sounds involve significant nonlinear effects. One example is the interaction between two objects that are in collision. This interaction may be short compared to the total duration of the resulting impact sounds, however, the interaction is nonlinear and can significantly impact the attack of the sounds. Another example is acceleration noise produced when an object is rapidly accelerated through air. Acceleration noise is nonlinear and perceptually significant for small objects such as shards of broken glass or ceramic objects (Chadwick et al., 2012). Some prior work has attempted to model the nonlinear residual components of impact sounds (Ren et al., 2013), though further analysis of these effects may improve analysis of real-world sounds and synthesis of virtual sounds.

Similarly, rigid objects are only one category of objects that may appear in environments and produce sound through interaction. The proposed methods will struggle to virtualize deformable objects, thin-shell objects, and objects with heterogeneous material. While objects such as mugs and bottles are common examples of rigid objects that produce modal sound, the proposed methods will also struggle when these objects contain liquid, altering their impact sound (Wilson et al., 2017). Recent methods propose generalizable wave-based frameworks for simulating a wider variety of physical sounds, though these methods are time-consuming and limited to offline synthesis (Wang et al., 2018).

Finally, the presented studies have found that no current damping model optimally represents all rigid-object materials. The damping models proposed in this dissertation may provide more accurate modeling of a subset of materials, but in that case it is still a challenge to identify which damping model to apply for a given real-world material. However, our methods for damping parameter estimation are designed to be easily extensible to estimate material damping parameters for future damping models.
1.3 Future Work

In the future, I would like to address some of these limitations by moving beyond linear models for modal sound. One of the primary challenges here would be maintaining sufficient runtime performance. Sound synthesis in particular must run in real-time to be applicable to interactive virtual environments. Parameter estimation methods do not have real-time requirements, but performance is still important in order to be practically useful for object virtualization.

The proposed probabilistic damping model (??) greatly improves the robustness of damping parameter estimation in the presence of confounding factors. Similar probabilistic approaches could be applied to other tasks. One example is estimation of object parameters relating to rigid-body dynamics, such as the coefficients of friction and restitution. Beyond object modeling, probabilistic methods may be applicable for estimation of liquid properties (e.g. viscosity) or deformable object properties (e.g. stiffness).

Alternatively, learning-based methods may provide further improvement in multiple areas. Learning-based methods have not yet been applied to the task of material parameter identification, and it remains to be seen how their results would compare against current methods. Sound synthesis is primarily still performed through physical simulation, but there have been recent advances in generative neural network design. Work such as “Visual to Sound” (Zhou et al., 2017) and “Visually Indicated Sound” (Owens et al., 2016) suggest that data-driven approaches may be able to produce high-quality synthesized impact sounds.

My work has begun to explore object reconstruction from multimodal inputs, but current methods have been limited to producing classifications pertaining to object geometry. In the future, I aim to perform automatic audio-visual reconstruction of scenes possibly containing multiple objects. The first challenge is performing full object reconstruction, producing the complete 3D geometry of a novel object while leveraging audio-visual inputs. The second challenge is reconstructing multiple objects in the same scene—differentiating and segmenting them from one another. However, being able to virtualize an entire scene would greatly extend the applications of these techniques.

Finally, I would like to consider applications of my work to augmented reality settings. The user studies presented in this dissertation all focus on virtual settings, and it remains to be seen how the results would translate to augmented reality. Object virtualization may play a large role: in augmented reality it may be useful to “copy and paste” a real-world object. Sensory cohesion may be more difficult to maintain, as any virtual objects will be directly contrasted against real-world objects present in the scene.


