CHAPTER 1: SUMMARY AND CONCLUSIONS

My thesis statement states that "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 sensory conflict." To improve the expressiveness of object material models, I presented methods for performing modal sound synthesis with Generalized Proportional Damping and automatically estimating material damping parameters for any GPD-derived damping model from a single impact sound (**??**). To account for real-world phenomena that impact estimation of material damping parameters, I presented a method for automatic material parameter estimation using a probabilistic damping model that encodes these phenomena (**??**). To reduce sensory conflict during interaction with virtual objects, I presented methods for multimodal interaction with textured surfaces using unified texture representations of detail (**??**). To further reduce sensory conflict, I presented a method for estimating object shape and material using joint audio and visual input (**??**). In this chapter, I will summarize these topics.

1.1 Summary of Results

In this section, I summarize the results found by myself and my collaborators. Modal sound synthesis simulates impact sounds produced when a rigid object is struck by modeling the object's vibrations. However, it is limited by the Rayleigh damping model, which is a linear approximation to a more complex phenomenon. Our method for deriving additional damping models for modal sound synthesis increases the expressiveness of material models by better capturing nonlinear damping behavior that Rayleigh damping would only approximate. Our single-sound damping parameter estimation method works for all damping models, making it easy to create virtual objects using these models. In our perceptual study, no damping model was consistently superior, demonstrating that Rayleigh damping cannot express the full variability in damping behavior.

This single-sound damping parameter estimation method is limited in that it requires significant knowledge of the object in addition to the recorded sound, and that the sound must be recorded in a carefullycontrolled environment. These limitations are addressed by our probabilistic damping model, which models real-world phenomena that influence damping estimates and enables robust damping parameter estimation. The only inputs are impact sounds recorded in less-controlled environments, making preparation very simple. Our human hand-tuning study establishes a baseline for human performance on the damping parameter estimation task, which future automated methods may compare against. Our perceptual evaluation found that sounds synthesized using our automatically-estimated parameters produce a pattern of errors similar to that of sounds synthesized using hand-tuned parameters, indicating high perceptual similarity with less human effort. 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.

These methods for damping parameter estimation focus on the perceptual realism of auditory object interactions, but do not account for object interaction's inherently multimodal nature. Our method for multimodal interaction with textured surfaces focuses on the perceptual realism of object interaction through reduced sensory conflict. Our texture identification study found that users perceived texture identification to be easiest 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 can guide the design of interactions with textured surfaces, and suggest that multimodal interaction using unified representations of detail can reduce sensory conflict.

Damping parameter estimation methods are also limited by operating in isolation from visual object understanding methods. Our Impact Sound Neural Networks reduce multimodal sensory conflict by using joint multimodal inputs. ISNN-A (audio-only input) and ISNN-AV (combined audio-visual input) provide accurate identification and classification of object shapes and material. ISNN networks are particularly useful when estimating properties of transparent or occluded objects. Our ISNN-A network outperforms models such as SoundNet (Aytar et al., 2016) on audio-only object identification tasks, while our multimodal ISNN-AV network outperforms the visual-only VoxNet (Maturana and Scherer, 2015) on the ModelNet dataset. These results suggest that joint audio-visual estimation of object properties can improve multimodal interaction with virtual objects created based on those properties.

1.2 Limitations

My proposed methods improve both the expressiveness of object damping modeling and multimodal interaction with virtualized objects, but some limitations remain. While the limitations of each method are discussed in its respective chapter, this section relates to the general methodology and assumptions made across my work.

1.2.1 Rigid Object Modeling

First, my proposed methods use linear models of object vibrations. While linear models are critical for real-time synthesis, impact sounds involve significant nonlinear effects. One example is the nonlinear interaction between two object in collision. This interaction is short compared to the total duration of the resulting impact sounds, however, it can significantly affect 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 (Chadwick et al., 2012). Prior work has attempted to model these nonlinear components of impact sounds as a residual (Ren et al., 2013), though further analysis of these effects may improve understanding of real-world sounds and synthesis of virtual sounds.

Current damping models are also limited. My presented studies found that no current damping model optimally represents all rigid-object materials **??**. The damping models proposed in this dissertation provide more accurate modeling of a subset of materials, but it is a challenge to identify which damping model is ideal for a given real-world material. My methods for damping parameter estimation (**????**) easily extend to estimate material damping parameters for future damping models.

1.2.2 Object Virtualization

To create realistic virtual versions of real-world objects (object virtualization), virtualizing only an object's audio-material is insufficient. In my work, I often neglect that an object's shape, surface appearance, tactile roughness, and even smell are important attributes of a virtualized object. My parameter estimation methods produce audio-material parameters which directly translate to a virtual object, but must rely on other methods for virtualization of any other object attributes. The ISNN networks take a step towards a more unified multimodal approach to virtualization, but sacrifice the ability to estimate quantitative material parameters.

I have focused on virtualization of rigid objects, but that is only one category of objects that may be expected to produce sound in virtual environments. My proposed methods will struggle to virtualize deformable objects, thin-shell objects, and objects with heterogeneous material—mugs and bottles are rigid objects, but produce different sounds depending on how much liquid they contain (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 would need to be significantly accelerated for interactive sound synthesis (Wang et al., 2018).

1.2.3 Multimodal Interaction

Without a complete multimodal object virtualization pipeline, virtualized objects are not ideal recreations of their real-world analogues. Many of the objects virtualized in this dissertation were carefully measured by hand to manually construct a 3D model, and surface textures were often selected from existing datasets as approximate matches. These virtual objects may have had accurate virtualized audio-material parameters from real-world objects, but the properties that had been picked by hand may have caused sensory conflict. Some of my perceptual studies (but not all) had subjects performing multimodal interaction with virtual objects; these studies may have been impacted by this sensory conflict.

Furthermore, interaction may be limited by hardware and software constraints. For hardware, there are two devices for object interaction across this dissertation, both with limitations. First, the PHANToM haptic device provides force feedback but covers a small working area, limiting object size. Second, the HTC Vive greatly expands the working area to the size of a small room, but only simulate haptics with vibrations. For software, many demos were implemented in game engines, which often sacrifice physical accuracy for computation speed, affecting rigid-body collision dynamics and surface appearances. With these and other technical constraints, even an object that has been perfectly virtualized may still not reproduce the same interactions in a virtual environment.

1.3 Future Work

In this section, I discuss five research directions for future work. These are: sound synthesis with nonlinear models, probabilistic modeling for other physical phenomena, learning-based methods for sound synthesis, automatic audio-visual object reconstruction, and extension of my methods to an augmented reality setting.

To improve the realism of synthesized impact sounds, one direction of research would be to use *nonlinear* models for object vibrations. Nonlinear models could more accurately simulate complex modes and nonlinearities during object collisions, leading to richer attack at the start of synthesized impact sounds. One of the primary challenges would be maintaining sufficient runtime performance. A sound synthesis module must provide samples at 44 kHz to be applicable to interactive virtual environments, but nonlinear methods tend to be computationally intensive. Parameter estimation methods would also need to be adapted for nonlinear models, and although they do not have real-time requirements, performance is still important to be practically useful for object virtualization.

Probabilistic models such as the one described in **??** can improve the robustness of estimation tasks in the presence of error or confounding factors. One application is estimation of object parameters relating to rigid-body dynamics, such as the coefficients of friction and restitution. More applications are estimation of liquid parameters (*e.g.*, viscosity) or deformable object parameters (*e.g.*, stiffness). Many optimization methods for these parameters seek to minimize a least-squares metric (Yang et al., 2016), which implicitly assumes error is normally distributed. If error can be more accurately modeled using a different statistical distribution, then a probabilistic maximum likelihood method may be ideal.

Learning-based methods may provide further improvement through two options. One option is applying learning-based methods to the task of material parameter identification, though it remains to be seen how their results would compare against current methods. The ISNN networks may serve as a starting point if their outputs can be adapted to produce quantitative parameters rather than classifications. The other option is performing sound synthesis directly through a learned generative model. 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., 2018) and "Visually Indicated Sound" (Owens et al., 2016) suggest that data-driven approaches may be able to directly synthesize high-quality impact sounds. One current challenge to address is the lack of an ideal dataset for rigid-object impact sounds; the Greatest

Hits dataset (Owens et al., 2016) covers a breadth of sounds but does not have the depth for a method focused on rigid objects.

Virtualization would ideally be performed through automatic audio-visual *reconstruction* of scenes possibly containing multiple objects. My methods provide a step in this direction, but are limited to producing *classifications* pertaining to object geometry. The first challenge is performing full object reconstruction: producing the complete 3D geometry of a novel object while leveraging audio-visual inputs. Reconstruction from audio alone is an underconstrained problem (Kac, 1966), but using vision as a regularizer may allow complete and accurate reconstructions. 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 methods.

Finally, there are unrealized applications of this work to augmented reality (AR) settings. Augmented reality requires more understanding of the real world than virtual reality does, so the ability to estimate properties of real-world objects is key. If a real-world object near an AR user can be virtualized on the fly, it opens multiple possibilities. The virtualized object may be duplicated to other locations or visually modified for interior design visualization (Choo and Phan, 2010), and virtual agents could produce realistic interactions with the original object. One challenge in adapting my methods to AR is that there are known to be differences in human perception between the real world, virtual reality, and augmented reality (Jones et al., 2008). 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. Another challenge is that sensory conflict may be more difficult to avoid, as any virtual objects will be directly contrasted against real-world objects present in the scene.

BIBLIOGRAPHY

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.

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).

Choo, S. Y. and Phan, V. T. (2010). Interior design in augmented reality environment. *International Journal of Computer Applications*, 5(5):16–21. Published By Foundation of Computer Science.

Jones, A., Swan, J. E., Singh, G., and Kolstad, E. (2008). The effects of virtual reality, augmented reality, and motion parallax on egocentric depth perception. In *2008 IEEE Virtual Reality Conference*, pages 267–268.

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

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.

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

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.

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.

Wilson, J., Sterling, A., Rewkowski, N., and Lin, M. C. (2017). Glass half full: sound synthesis for fluid–structure coupling using added mass operator. *The Visual Computer*, 33(6):1039–1048.

Yang, S., Jojic, V., Lian, J., Chen, R., Zhu, H., and Lin, M. C. (2016). Classification of prostate cancer grades and t-stages based on tissue elasticity using medical image analysis. In Ourselin, S., Joskowicz, L., Sabuncu, M. R., Unal, G., and Wells, W., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 627–635, Cham. Springer International Publishing.

Zhou, Y., Wang, Z., Fang, C., Bui, T., and Berg, T. L. (2018). Visual to sound: Generating natural sound for videos in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.