CHAPTER 1: INTRODUCTION

Sights and sound are all around us, in both real and virtual worlds. At times, it is useful to unmute our speakers or put on earphones so we can not only see what looks real but also hear how it sounds. By modeling how objects and fluids should behave and sound, we can generate impact sounds of a user striking an object, colliding sounds of objects interacting with one another, and other object and environment sounds. While visual information provides a great amount of context for what one can expect, auditory cues can assist with complementary data; for example, by differentiating between visually similar materials or the type and amount of liquid in an opaque container. Audio can also provide primary data when vision is unavailable; for example, unlit scenes or sounds outside the current field of view. Whether modeling by physically based phenomena or training a multimodal neural network, both audio and visual information play an important role in learning and processing of data for scene and object understanding.

Not only can additional modes of data provide more information, they can also reaffirm one another. For example, if we only read the captions of a video, we may miss context of the scene or speakers. Information from auditory cues may be lost. Furthermore, we can also use the sound to verify the visual and textual data based on the synchronization between modes. We learn this through experience and understanding. The uncanny valley effect is also described in terms of visual resemblance; however, the same can be applied to our perception of sound, though likely not as pronounced. The effect of audio and visual together matters as well. Individual modes and their interplay represent research areas in multimodal learning, cross-modal self supervision, and transfer learning, to name a few. This dissertation contributes to the simulation of sound for rigid body objects, with and without liquid, and uses both real and synthetic audio as an additional mode to visual data for learning and processing tasks.

1.1 Motivation

> *‘ The screen is a window through which one sees a virtual world. The challenge is to make that world look real, act real, sound real, feel real. ’* (Sutherland, 1965)

Virtual environments: as the excitement and economic potential for interactive virtual reality (VR) and augmented reality (AR) increases, modeling of the physical world is imperative to realism of an immersive experience, a sense of being there. In addition to sight, sound is also integral to the level of immersion and sense of presence in virtual and augmented reality (Cummings and Bailenson, 2015). This interest has motivated prior work in 3D sound synthesis, particularly in real-time. Audio has been used to guide user attention and highlight parts of a scene outside current field of view. In the case of redirected walking, sound can also serve as a distraction. The virtual scene can be manipulated in such a way that the user can travel through a virtual world that is larger than the physical working space without the user noticing. Whether or not audio and visual information are processed together in a single pipeline or separately, the presentation of the modes should be synergistic to prevent a distraction from any of the senses which can cause a 'break in presence' (Sanchez-Vives and Slater, 2005).

Sound synthesis: traditionally, sounds have been added post-production by Foley artists who recreate sounds for film and other media. However, today's virtual environments expect real-time interactivity. Therefore, game engines and VR systems are incorporating physically-based graphics and sound simulation algorithms for interactive and realistic effects to help users remain immersed in the experience. This modeling can be done based on the vibration and deformation of an object from interacting with another surface. This deformation generates pressure waves through a medium such as air or a liquid. Our ears hear the variation in pressure as sound, although we may not see the vibrations. This harmonic motion can be modeled as an underdamped spring mass system.

$$mx'' + dx' + kx = 0 \tag{1.1}$$

where m is mass, d is damping, k is stiffness, and x is displacement. Sound synthesis and physically-based sound synthesis for rigid bodies as well as liquids have been studied. A few other major categories include fractures, fire, and thin shell.

Since both sound and graphics can be physics-based, the graphics pipeline can be naturally extended to generate sound. Humans can hear frequencies from between 20 Hz to about 22 kHz, requiring applications to sample at a rate of 44 kHz based on the Nyquist Theorem, doubling the frequency we can sense. Rigid body sound has been modeled using modal analysis to decouple the problem into n independent, damped

vibration equations. By performing modal analysis to precompute frequency and damping, real-time sound synthesis can be achieved (O'Brien et al., 2002; Ren et al., 2013a; van den Doel et al., 2001). This is important such that the audio and visual information rendered from interactions of virtual objects with other objects, liquids, and the user reflect the current state of the virtual environment. Precomputing features, clustering sources, decoupling equations, and/or simplifying the computational model are techniques that have been used to achieve real-time performance and are explored in contributions to this dissertation.

Fluid-structure interactions: the dynamic sound synthesis model may be generalized to any object represented by a tetrahedral mesh and fluid by Lagrangian particles. The next extension of the pipeline is to account for these objects that contained a liquid. While each perform in real-time separately, the coupling of fluid-structure interactions could be compute intensive and needed to meet no-penetration and no-slip conditions.

$$(\frac{\partial u}{\partial t} - v) \cdot n = 0 \tag{1.2}$$

The no-penetration condition (Equation 1.1) occurs at the fluid-structure boundary where u is the deformation of the solid, v is the velocity of the fluid, and n is the outward normal.

$$(\frac{\partial u}{\partial t} - v) \times n = 0 \tag{1.3}$$

The no-slip condition (Equation 1.1) holds that the tangential velocity components have to be equal. If both independent boundary conditions hold, we have $\frac{\partial u}{\partial t} = v$. The problem can be simplified to a single system, sometimes referred to as a rigid doubly body, by assuming:

1. Solid object is impermeable
2. Fluid is incompressible
3. Fluid motion coincides with structure motion (also known as a non-moving domain)
4. Fluid is at rest in hydrostatic equlibrium
5. Fluid is inviscid

Applying Newton's third law, for every action, there is an equal and opposite reaction, the resting forces along the boundary of the fluid-structure interface can be calculated and included as an added mass to the sound dynamics equation, extending the sound synthesis pipeline to objects containing a liquid.

Multimodal learning: learning-based methods often primarily rely on visual feedback and human interaction; for example, state-of-the-art vision-based techniques for image classification (Deng et al., 2009) and object detection (Liu et al., 2016; Redmon et al., 2016; Ren et al., 2015a) in images and video, to name a few. Prior work for the liquid pouring task and object detection in robotics have also used visual sensing for volume estimation and tracking. With many of these methods using video as an input, multimodal learning with both audio and visual data can improve processing and performance. Fused modalities also cover edge cases that can be a challenge for a single model. For visual data, noise from blur, poor illumination, or occlusions can cause error. On the other hand, environmental noise, varying room acoustics, or mixed audio from other sound sources can prove difficult with audio inputs.

Natural Language Processing (NLP) has demonstrated the use of multimodal learning for visual question and answering systems (Fukui et al., 2016; Ilievski and Feng, 2017), video captioning (Pasunuru and Bansal, 2017; Wang et al., 2018). Audio-visual have also been used for speech separation (Zhao et al., 2018) and object classification (Anusha and Roy, 2015). Rather than using extra sensors such as contact microphones (Clarke et al., 2018) or thermal imaging cameras (Schenck and Fox, 2017), frames of audio and visual data are used from the recorded video. Various techniques have been used to fuse these multiple modes of data, such as merge layers of concatenation, add, and multiply to combine separate input streams. Bilinear modeling has also been used to learn multiplicative interactions of differing input types (Gao et al., 2015; Yu et al., 2017b; Park et al., 2016). A simple multi-modal bilinear model can be represented:

$$z_i = x^T W_i y \tag{1.4}$$

where x and y are mode features, $W_i \in R^{mxn}$ is a projection matrix, and z is the output blinear model.

3D reconstruction: a number of algorithms exist to generate 3D shape from 2D and other sensory information. Passive methods use sensors (e.g. camera) to capture details (e.g. RGB) about an object for reconstruction without any interference or projections into the scene. On the other hand, active reconstruction techniques (e.g. RGB-D) use infrared projectors to illuminate and detectors to measure the radiance on the object's surface. Using commodity sensors such as the Microsoft Kinect and GPU hardware allow

for both static (Golodetz* et al., 2015; Izadi et al., 2011) and dynamic (Dai et al., 2017b; Newcombe et al., 2015) scenes to be scanned in real-time. 3D scene reconstructions have also used sound such as time of flight sensing (Crocco et al., 2016).

Results of reconstruction of 3D geometries can also serve inputs to other learning based algorithms. For instance, 3D points have been used as inputs to train neural networks for other downstream tasks such as object classification, segmentation, and tracking (Qi et al., 2016a). Reconstruction research has generated large amounts of 3D scene (Silberman et al., 2012a; Song et al., 2017) and object (Lai et al., 2011; Singh et al., 2014; Wu et al., 2015b) data that can be used for training vision-based neural networks for classification, segmentation, and other downstream tasks.

1.2 Scope of this dissertation

There are a number of training datasets, neural network architectures, technologies, and active research areas for multimodal learning, especially in the area of audio and visual data from video. Applications in these areas range from Virtual and Augmented Reality, e.g., sound synthesis, reconstruction, inference, etc. to expanding methods to handle a wider variety of surfaces and scenes, such as illumination, reflectivity, texture, and occlusion. This dissertation focuses on coupling fluid-structure (chapter 3) and audio-visual classification (chapter 4), tracking (chapter 5), and reconstruction (chapter 6 and chapter 7) with demonstrations in multimodal learning and virtual reality.

1.3 Thesis Statement

My thesis statement is as follows:

Coupling multimodal information enhances task performance and processing of audio-visual learning based methods for fluid-structure sound synthesis, liquid pouring sequences, object tracking, and 3D reconstructions while also allowing for single mode application for special cases.

To support this thesis, I present four methods; one method to efficiently synthesis sound of objects containing a liquid, two methods to accurately estimate liquid pouring sequences and track objects using audio-visual neural networks, and one method to use audio to enhance scene and object reconstructions using mobile devices.

1.4 Main Results

1.4.1 Sound Synthesis for Fluid-Structure Coupling

Previous sound synthesis research has focused on single systems only, either solid or liquid but not both. Since not all single mode, sound simulations achieve real-time performance, modeling the variation in sound from a coupled vibrating fluid-structure system could be computationally expensive. This work was the first to synthesize sound for a system containing both a rigid body object and liquid (referred to as a fluid-structure coupling).

In chapter 3, I present a fast and practical method for simulating the sound of rigid body objects that contain liquid. To maintain real-time, interactive performance, we modify the existing modal synthesis pipeline by adding pre-processing steps. Those steps are to identify mesh nodes of the object that bound the liquid and to then modify the mass matrix of the structural object by an amount proportional to the liquid density and volume.

The main contributions of my work are:

1. Transforming the problem into a single fluid-structure system using the *added mass operator*;
2. Enhancing the sound synthesis pipeline with pre-processing steps for objects containing a liquid;
3. demonstrating the proposed method in interactive 3D VR applications.

Actual recordings versus synthesized frequencies were compared for varying amounts of liquid and results were less than 5% relative error. The interactivity of the algorithm was demonstrated with VR applications of a simulated liquid xylophone and kitchen scene of different containers, liquids, and volumes.

1.4.2 Analyzing Liquid Pouring Sequences via Audio-Visual Neural Networks

Prior work to estimate liquid poured amounts often require predefined amounts in the source container or rely on visual data. To compensate for vision-based challenges such as occlusion and transparency, this work uses audio from the pouring sequence to augment audio and visual only methods.

In chapter 4, I introduce audio and audio-visual neural networks in the form of multimodal convolutional neural networks (CNNs) to perform weight estimation, overflow detection, and content and container classification for robots pouring liquids.

The main contributions are:

1. Training, validation, and test data generated from audio recordings and video images with ground truth measurements from a digital scale
2. Audio-based CNN for multiclass weight estimation and binary classification for overflow detection by robotic systems
3. Audio-augmented neural network enhancing the audio only based method with fused visual inputs for robots pouring contents into various target containers
4. Pouring content and target container classification for robots, based on pouring sequence audio data

Upto 91.5% of the audio intervals for the robot pouring sequences were classified within 0.4 oz using audio-visual data. This resulted in an average error of 0.2 oz. The sound from pouring the liquid was also used to predict the type of liquid and target container.

1.4.3 Audio-Visual Object Tracking of Multiple Objects

Visually based object trackers can run into challenges when object collide, occlude, or appear similar but differ in material. By using audio of the impact sounds from object collisions, rolling, etc., an audio-based technique may be used in conjunction with other neural networks to augment visually based object detection and tracking methods. In chapter chapter 5, I describe the first use of an audio-visual neural network for tracking tabletop sized objects and enhancing visual object trackers.

The key contributions of this work include:

1. An end-to-end, jointly trained audio-visual object tracker (AVOT) to enhance visual object tracking
2. Ground truth bounding boxes for virtual scenes from the Sound-20K audio-visual dataset with 1, 2, and 3 objects
3. Scheduler for object detection re-initialization based on audio onset detection when using multi-modal tracking

By fusing audio with visual data, the audio-visual object tracker (AVOT) achieves upto 78% intersection over union (IoU) post-collision tracking accuracy compared to 69% using state-of-the-art deep learning visual trackers.

1.4.4 Audio-Augmented Scene and Object Reconstruction

In chapter 6, I introduce echoreconstruction, an audio-visual method that uses reflecting sounds to aid in geometry and audio reconstruction. Scenes containing open and reflective surfaces often lead to existing techniques reconstructing objects behind (in the case of transparent glass) or in front of (in the case of mirrors) the object. By using pulsed audio from a mobile device, inferences from a convolutional network can detect and estimate depth to the reflecting surface. Key results include:

1. EchoCNN, a fused audio-visual CNN architecture for classifying open/closed surfaces, their depth, and material

2. EchoReconstruction, a staged audio-visual 3D reconstruction pipeline that uses mobile phones to enhance scene geometry containing windows, mirrors, and open surfaces with depth filtering and inpainting based on EchoCNN inferences

3. Semantic rendering of window and mirror in audio-augmented reconstructions based on point of view (e.g. environment outside of the window or reflected view of a TV)

4. Real and synthetic audio-visual ground truth data for multiple scenes containing windows and mirrors in addition to reflection separation data (direct, early, or late reverberations)

Overall, 71.2% of hold out reflecting sounds were correctly classified as an open or closed boundary and 71.8% of 1 second audio frames were correctly classified as 1 ft, 2 ft, or 3 ft away from the surface based on audio alone; 89.5% when fused with its corresponding image. Pulsed sounds were emitted a maximum of 3 feet away to remain in the free field. Beyond that, there will be less noise reduction due to reflecting sounds in the reverberant field (Egan, 1988).

In chapter 7, I detail a multimodal single and multi-frame LSTM autoencoder for 3D reconstruction using audio-visual input. Existing methods may experience difficulties in cluttered environments with multiple objects causing occlusion. To address such limitations, the method adds audio as another input, specifically *impact sounds* resulting from object to object or scene interactions. The main contributions of this work can be summarized as:

1. A multimodal LSTM autoencoder neural network for geometry and material reconstruction from audio and visual data

2. The resulting implementation has been tested on voxel, audio, and image datasets of objects over a range of different geometries and materials

3. Experimental results of our approach demonstrate the reconstruction of single sounding objects and multiple colliding objects in a virtual scene

4. Audio-augmented datasets with ground truth object tracking bounding boxes

Single view ShapeNet resulted in IoU metrics of 21.2% for audio and 32.6% for audio-visual. 10 Sound20K views resulted in 37.15% and 69.8% IoU for audio and audio-visual respectively.

1.5 Contributions of this dissertation

For sound synthesis: (1) Transforming the problem into a single fluid-structure system using the *added mass operator*. (2) Enhancing the rigid-body sound synthesis pipeline with pre-processing steps for objects containing a liquid. (3) Demonstrating the proposed method in interactive 3D VR applications.

For analyzing pouring sequences: (1) Training, validation, and test data generated from audio recordings and video images with ground truth measurements from a digital scale. (2) Audio-based convolutional neural network for multi-class weight estimation and binary classification for overflow detection by robotic systems. (3) Audio-augmented neural network enhancing the audio only based method with fused visual inputs for robots pouring contents into various target containers. (4) Pouring content and target container classification for robots, based on pouring sequence audio data.

The broad contributions of this dissertation are new real-time fluid-structure coupling methods, new audio-visual classification and tracking, and prototype audio-augmented object and scene reconstruction on mobile devices.

1.6 Organization

The remainder of this dissertation is organized as follows. The discussion of fluid-structure coupling begins with the sound synthesis of objects containing a liquid using the added mass operator in chapter 3. This is followed by my work on analyzing liquid pouring sequences using audio-visual neural networks in chapter 4. Next, I cover my method for audio-visual object tracking in chapter 5. Finally, I discuss a model for using audio on mobile devices to enhance 3D reconstructions of scenes (chapter 6) and objects

(chapter 7). I conclude my dissertation in chapter 8 by presenting a summary of this work, its contributions, and a discussion of future work.