Inverting Audio-Visual Simulation for Shape and Material Perception

Zhoutong Zhang*[1], Jiajun Wu*[1], Qiujia Li[2], Zhengjia Huang[3], Joshua B. Tenenbaum[2] , William T. Freeman[1,4]

[1]Massachusetts Institute of Technology, [2]University of Cambridge, [3]ShanghaiTech University, [4]Google Research

1. Introduction

Humans perceive objects through both their visual appearance and the sounds they make. Given a short audio clip of objects interacting, humans can recover rich information about the materials, surface smoothness, and the quantity of objects involved [3]. Although visual information provides cues for some of these questions, others can only be assessed with sound. Figure 1 shows an example: objects with different masses and Young's moduli may have almost identical appearance, but they make different sounds when impacted, and vice versa.

(A) paper cup (B) ceramic cup (C) ceramic plate

Figure 1: Audio and visual data provide complementary information: visual cues tell us that A and B are cups and C is a plate; auditory cues inform us that A is made of a different material (paper) than B and C are (ceramic).

Since collecting large-scale audio recordings with rich object-level annotations is time-consuming and technically challenging, We introduce an alternative approach to overcome such difficulties: synthesizing audio-visual data for object perception. Our data synthesis framework is composed of three core generative models: a physics engine, a graphics engine, and an audio engine. The physics engine takes objects' shapes, material properties, and initial conditions as input, and then simulates their subsequent motions and collisions. The graphics engine renders videos based on the simulated object motion. The audio engine, built upon previous works [2], synthesize the audio using the output of physics engine.

*Equal Contribution

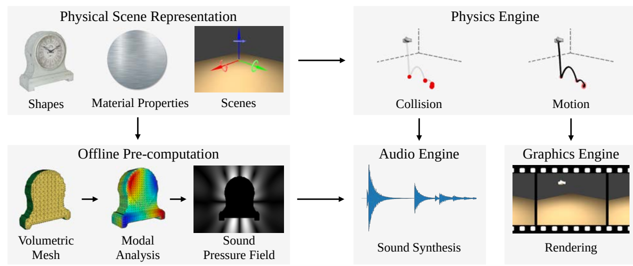


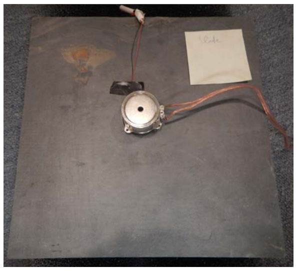
Figure 2: Our generative model for audio-visual scenes. Given object shapes, material properties, and a scene configuration, a physics engine simulates both object collisions and motion. An audio engine then takes the collision data and pre-computed object mode shapes for sound synthesis. Graphics engine renders accompanying video.

With our generative model, we built a new synthetic dataset, Sound-20K, with audio-visual information. We show, on both Sound-20K and real-world datasets, that visual and auditory information contribute complementarily to object perception tasks and further demonstrate that knowledge learned on our synthetic dataset can be transferred for object perception on two real-world video datasets, Physics 101 [5] and The Greatest Hits [4].

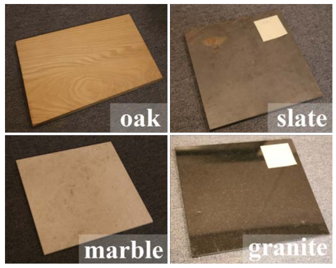
In addition, the audio and physics engine can perform in real time, which enables us to infer the latent variables that defines object shape, material properties and initial pose in an analysis-by-synthesis style. In short, given an audio clip, we aim to find a set of latent variables that could best reproduce it. We use Gibbs sampling over the latent variables and pass them to our synthesize engine. The likelihood function is given by the similarity between the input and output audio. We show that with simple similarity measure, such as l_2 distance over spectrogram, such inference scheme performs reasonably well.

2. Synthesis Engine

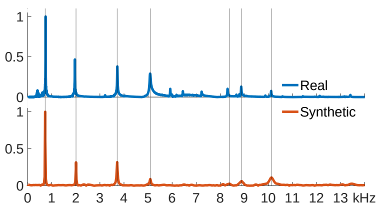
Our design of the generative model originates from the underlying mechanism of the physical world: when objects move in a scene, physical laws apply and physical events like collisions may occur. What we see is a video rendering



(a) The recording setup



(b) The four tiles we used



(c) Real *vs*. synthetic spectra of granite

Tiles	Labeled *real* (%)
Oak	45.3
Slate	50.5
Marble	52.6
Granite	46.3

(d) Behavioral results

Figure 3: We validate our audio synthesis pipeline through carefully-designed physics experiments. We record the sound of four tiles of different materials (a-b), and compare its spectrum with our synthesized audio (c) with corresponding physical properties. We also conducted behavioral studies, asking humans which of the two sounds match the image better. We show results in (d).

of the scene with respect to object appearance, lighting, *etc*., and what we hear is the vibrations of object shapes caused by physical events. Our generative model therefore consists of a physics engine at its core, an associated graphics engine and an audio engine, as shown in Figure 2. The generative model can be decomposed into an on-line stage and an off-line stage. The off-line stage computes object's acoustic properties, which is used be the on-line stage to synthesize audio in real-time.

3. Audio Synthesis Validation

We validated the accuracy of our audio synthesis by comparing it with real world recordings. We recorded the sounds made by striking four plates of different materials (granite, slate, oak and marble) as shown in Figure 3b. The audio was measured by exciting the center of the plates with a contact speaker and measuring the resulting vibrations with a piezo-electric contact microphone placed adjacent to the speaker (shown in Figure 3a). All measurements were made in a sound-proof booth to minimize background noise in the recording.

We validated the accuracy of our synthetic sounds by comparing the spectrum of synthetic audio with real recordings. Figure 3c shows the spectrum comparison between the synthetic sound and the real recording of the granite tile. We also designed a human perceptual study in which 95 people were asked to judge whether the recording or the synthetic was more realistic. Table 3d shows the percentage of people who labeled synthetic sounds as real.

4. Experiment Results

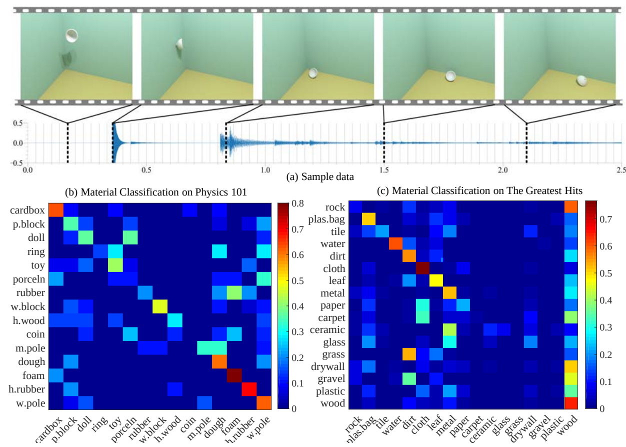


Figure 4: Sample data and material classification results

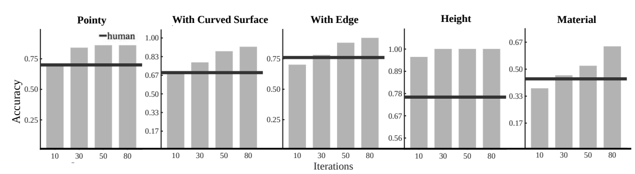


Figure 5: Human performance and Gibbs sampling result comparison. The horizontal line represents human performance for each task. Our algorithm closely matches human performance.

We first built Sound-20K using our generative model with 39 3D shapes sampled from ShapeNet [1]. Each shape can randomly chose from a pool of 7 material setup. We designed 22 scenarios with different level of complexity. Through randomly sampling objects, materials, scenarios and initial pose, we generated 20,378 audios of objects falling and interacting with corresponding videos. Samples are shown in Figure 4a.

We then use Sound-20K to train neural networks on martial classification and shape attribute recognition. Our results show that auditory information is complement to visual cues on such tasks. We further demonstrate knowledge learned on Sound-20K can be transferred to controlled real world scenarios in Physics 101 [5]. Results are shown in Figure 4b and Figure 4c.

Using the on-line stage of our generative model, we further explored the task of inferring the latent variables that generates a given audio. The latent variables consists of object shape, material properties and height of the drop. In this experiment, we use 14 primitive shapes with uniformly sampled physical properties and initial pose. Using Gibbs sampling over the latent variables, we aim to find the configuration that could best reconstruct the input audio. Our algorithm performs similarly as human subjects in tasks on inferring objects' shape and material properties, as shown in Figure 5.

References

[1] A. X. Chang et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[2] D. L. James, J. Barbič, and D. K. Pai. Precomputed acoustic transfer: output-sensitive, accurate sound generation for geometrically complex vibration sources. *ACM TOG*, 25(3):987–995, 2006. 1

[3] A. J. Kunkler-Peck and M. Turvey. Hearing shape. *Journal of Experimental psychology: human perception and performance*, 26(1):279, 2000. 1

[4] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *CVPR*, 2016. 1

[5] J. Wu, J. J. Lim, H. Zhang, J. B. Tenenbaum, and W. T. Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, 2016. 1, 2