Physics-based Audiovisual Simulation in Automatic Online Perception

Aaditya K. Singh
Department of Computer Science
Massachusetts Institute Of Technology
aaditya@mit.edu

1 Introduction

Two of the most fascinating things about the human mind are the human mind's ability to react so quickly, and its unique grasp of basic physical intuition, even from a very young age. In this project, we attempt to see how these two skills might intersect and how our intuitive physics intuition may be responsible for our quick reaction times.

Due to the latency between a planned movement in the brain and execution of the movement (which, although short, is still very present), some human reactions are almost impossibly fast. A possible explanation for this comes via the view of efference copies. It is commonly believed in the field that a copy of the signal to the motor cortext is actually sent to the sensory cortex to "prepare" it for possible sensations, and that this could lead to our fast reaction times. This view has been pushed forward in [2] and [3]. In [2], it is shown that externally-generated forces are perceived as stronger than self-generated forces. This could potentially be since our brain "expects" a force and somehow "prepares" for it when the force is self-generated. Similarly, in [3], efference copies are believed to underly the basic phenomena of tickling, and why humans can't tickle themselves. Thinking about it in another perspective, efference copies and the brain's "preparation" for the future could be viewed as simulation forward in time. Could our brain secretly be living a few milliseconds in the future and actually basing its decisions off of that future state to ensure good reaction times? To what extent does this simulation dictate our behavior?

One of the most fundamental behaviors of humans, as mentioned earlier, is our interaction with other objects in the world. Underlying this interaction is a strong intuitive grasp of the physics of object motion [1]. In [1], strong evidence is provided for the existence of an intuitive physics engine (IPE) in the brain that is responsible for physics simulation forward in time. By quickly running many forward simulations, the IPE would explain how we can look at a stack of dishes and assess the stability almost instantly. However, could the extent to which the IPE dictates our actions be even more significant. What if our quick reaction times are actually a result of our brain's IPE simulating the world around is forward in time slightly, and making movement judgements off of those? In this project, we hope to gain insight on this matter.

Testing a belief about the brain's forward simulation of physics is difficult since we have to find a way to query this internal state/existence of forward simulation. In this project, we use a behavioral experiment centered around loudness judgments. After observing how human's perform on this task, another important aspect is creating a computational model to better understand the human behavior. In this report, I present results from a preliminary behavioral experiment delving into the above questions, as well as some ideas for models of the behavior and their shortcomings.

2 The Experiment

The behavioral experiment used to query the brain's physics-based audiovisual simulation is centered on loudness. To query the presence of simulation, we present incongruent stimuli to the subject and

MIT 9.520/6.860: Statistical Learning Theory and Applications (Fall 2018).

look at the reactions in these "trick" cases. Although designing this experiment (and this project) started as a volunteer UROP/independent research under Ilker Yildirim, all the work presented in this report was done for this class and completed this semester.

2.1 Design

The behavioral experiment centered on loudness judgments after observing a scene of a wooden block falling on a surface. The block used had three different possible shapes (see Figure 1), and it fell on one of three possible surfaces (see Figure 2). We expect loudness to be mostly independent of object shape and almost completely dependent on the material (specifically, the hardness) of the ground surface. Specifically, since foam is softer than wood, which is softer than the ceramic plate, we expect loudness judgments to be softest on foam, louder on wood, and loudest on the ceramic plate. This prior was made clear via the instructions before the experiment and the understanding of these assumptions was tested with a short pre-quiz before the experiment.

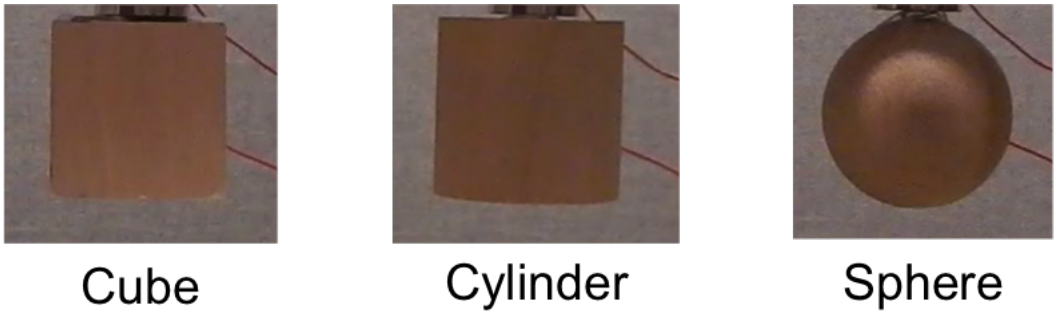

Figure 1: Three object shapes that were dropped

Figure 2: Three ground surfaces

The actual trials of the experiment consisted of videos starting with a freeze frame for 1 second. Then, the lights would "turn off" (the experimental setup would be hidden) and the block would drop onto the surface and a sound would be heard. Subjects were then asked to rate the loudness of a sound using a slider from 0 to 100 labeled "Very soft", "Medium", "Very loud". An example scene and transition can be scene in Figure 3. Note that in the first version of the experiment, this lights-off did not occur and subjects could see the bounce pattern exactly. However, this led to the incongruent stimuli standing out much more. The lights turning off adds some uncertainty to the prior that is anchored by the initial freeze frame.

The incongruent stimuli in the experiment, as you might guess, consisted of trials where the audio did not match the video. I encountered some issues in making this work (and spent a lot of time debugging/trying different approaches!). I was attempting to resynthesize the audio over the videos to make it seem like the surface was a different surface. The main issues were related to matching the complicated bounce patterns of the cube and cylinder to make the synthesized videos realistic and a rich stimuli set. Removal of synthesized artifacts in the audio was also difficult. In the end, I went with an approach using peak finding and extracting a "good" peak from a recording of a sphere dropping on a surface. Then, I use more peakfinding to extract the bounce times of a shape dropping on the surface. The synthesized audio is then creating by adding scaled versions of the "good" peak to a zero audio waveform. Note that the only scaling performed is the relative magnitude of bounces.

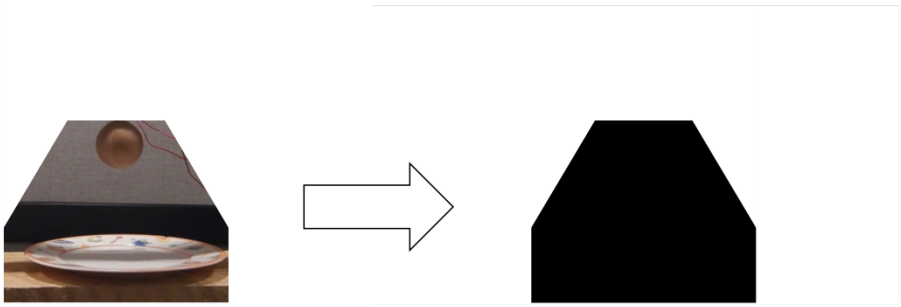

Figure 3: An example of the lights turning off in a scene

The magnitude of the first synthesized peak is always the same as the magnitude of the "good" peak. This normalizes against other possible variation in loudness, since all the audios in the experiment are actually just synthesized from 3 peaks (a sphere dropping on foam, wood, or ceramic). Furthermore, to normalize against any possible remaining synthesis artifacts, even the presented congruent stimuli were actualy sythesized audios (i.e. of a cylinder dropping on foam, synthesized using the foam "good" peak). For more details on this approach, please take a look at the code linked at the end of the report. MATLAB was used to create the synthesized audios, and ffmpeg was used for all other audio/video processing.

The actual experiment was run using psiturk. Although for this class I did not launch my full experiment on Amazon Mechanical Turk (since this is just the pilot experiment), I used psiturk so that, as I continue my research under Ilker, I can eventually launch on AMT to collect more data. For now, I just used the sandbox mode of psiturk and sent many of my friend groups the link to take the experiment. I had a total of 9 anonymized responses. In the experiment, each subject was first shown some instructional scenes, then asked to prove understanding of the priors with a short quiz (as mentioned above). Then, a practice scene was shown. After the instructions and practice scene were over, the subject had scene one example each of an object falling on each of the different surfaces. I did this to hopefully give subjects a way to compare the relative loudness and make judgments. The actual experiment consisted of 42 trials presented in a semi-randomized order. There was one trick trial of each possible incongruent combination (for a total of 3*2=6). The shape used was randomized for these trick trials. For the standard trials, 4 trials of each shape falling on each surface were presented (for a total of 4*(3*3)=36). The order was randomized such that no two trick trials come directly after one another, and the first four trials of the experiment are always congruent stimuli.

2.2 Results and Trends

The raw loudness judgments are presented in Figure 4. As expected, each participants responses appear to form 3 clusters (based on the audio). This is emphasized by sorting the audio into rows. Also, as a reminder, there are actually only 3 different loudnesses that are heard so a perfect observer would only have 3 dark dots, one in every row.

To look at the data as a whole (instead of on a subject by subject basis), some normalization must be performed. Whereas z-scores are a standard way of normalizing behavioral data, I used a slightly different approach. Since each subject's responses form three clusters, instead of z-scoring all the values, I separated the three clusters and z-scored them individually. Then, I could compare the z-scores between different participants ranking ceramic sounds. Specifically, I believe this is the easiest way to see if there is a general trend in how the participants judge the incongruent stimuli (i.e. for a ceramic sound but a foam scene) relative to the standard stimuli (i.e. a ceramic sound for a ceramic plate scene). These normalized results are presented in Figure 5.

In some of the subject responses, some interesting trends are observed. For example, subject 5 shows clear evidence on the two ends of the spectrum of the predicted outcome. When a block appears to fall on ceramic, but then lands on foam, the perceived loudness is way softer, perhaps because the body somehow "prepared" for the ceramic sound, which is typically loud. This effect is less pronounced for the wood sound since the wood sound is less loud. Similarly, when a block appears

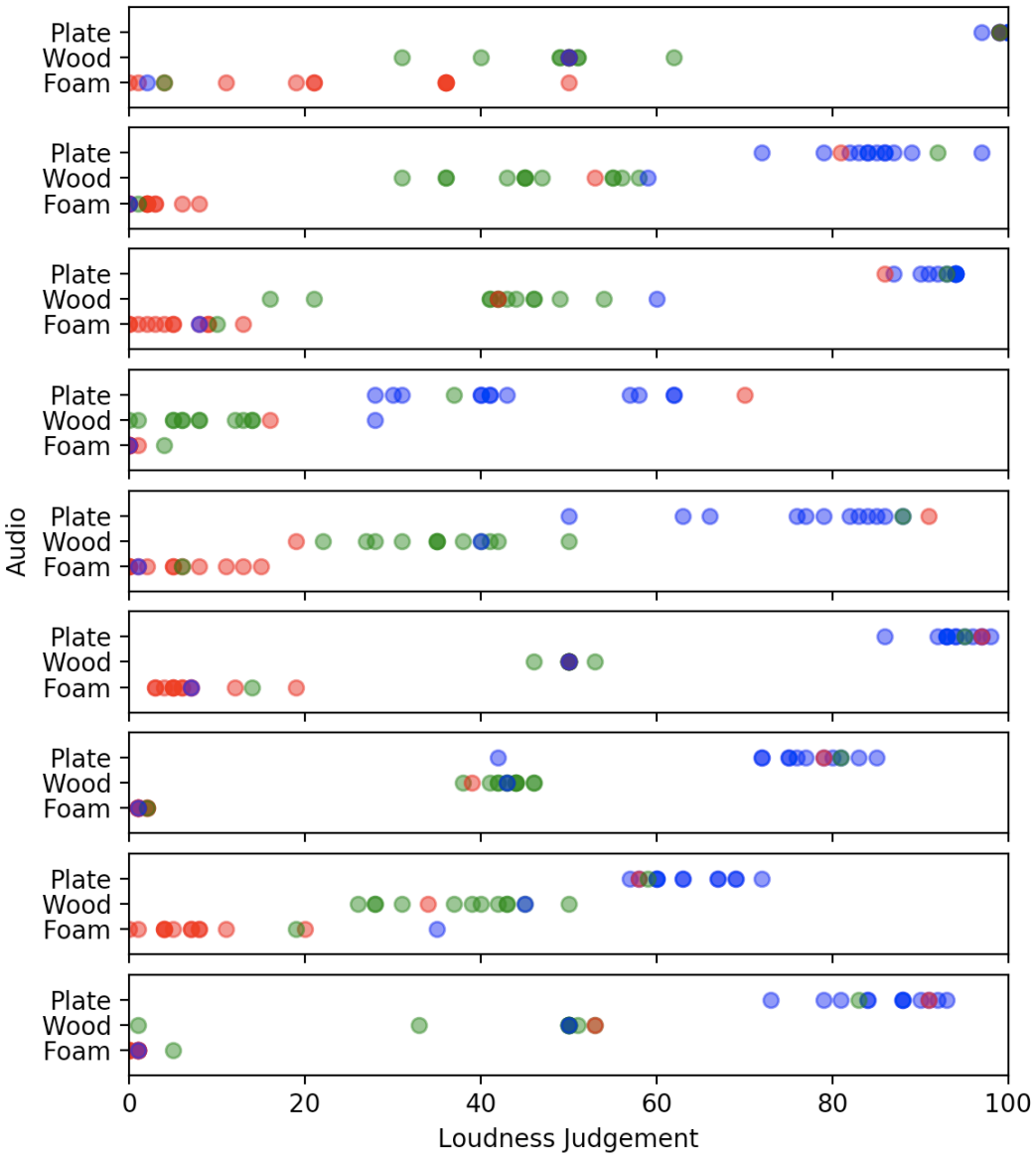

Figure 4: Raw results for the 9 participants. Each row in each plot represents a different audio heard. The color of the dot indicates the surface scene for that audio (Red = Foam, Green = Wood, Blue = Ceramic). Note that the dots are somewhat transparent so that overlapping areas appear darker.

to fall on foam but then lands on ceramic, the perceived loudness is way louder. Again, this effect is less pronounced if the object appears to fall on wood.

Despite promising results such as these, from Figure 5, we see that there is no clear overall trend. I have some possible theories for why this might be the case. First of all, some subjects, in the post-questionairre, indicated that they "zoned out" and were just listening to the sounds. This, of course, ruins the purpose of the experiment. Secondly, the slider from 0 to 100 may be adding too much noise to the measurements.

2.3 Steps forward

To combat the above issues, I plan on iterating on my epxeriment in the future. I think that the evidence of the predicted behaviors is present (as discussed above), but there are some aspects of the

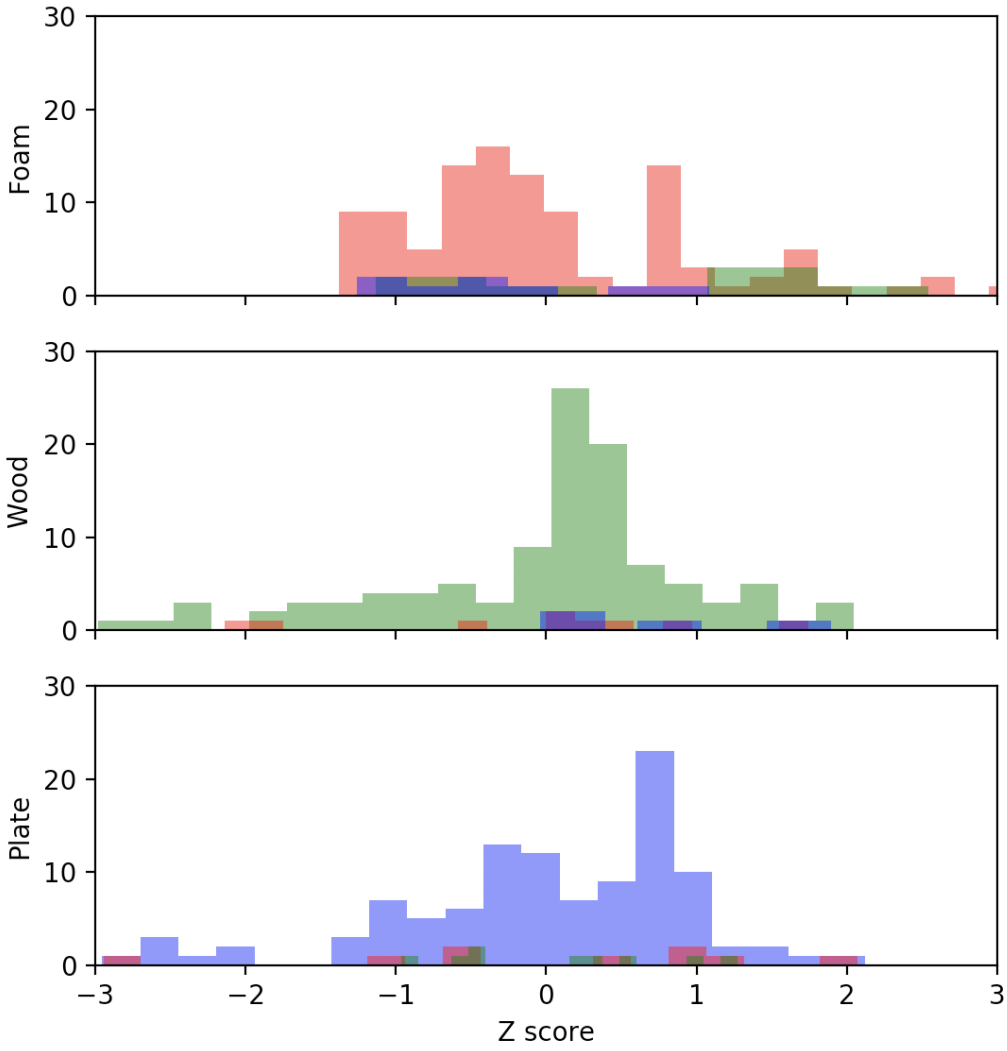

Figure 5: Normalized results for the 9 participants. Each subplot shows the z-score distribution for a different audio (see y-axis labels). The colors are the same as Figure 4. The incongruent stimuli distributions are also shown overlayed in the respective colors. There appears to be no clear pattern in the responses to incongruent stimuli.

experiment that could be improved upon. For example, the experiment could be made more engaging so that people pay attention to the visual cues (although I suspect this may also become more the case when there is monetary incentive via AMT). To combat the noise from the slider, I plan on conducting another pilot using a scale from 1 to 7 which will be much more discrete. My worries with this approach (and reasoning for not initially taking it) is that I fear it is not fine grated enough to register the slight difference we are aiming for with the incongruent stimuli. Finally, I think just collecting more data from more subjects could be fruitful in discovering larger trends.

3 Models

In this section, I discuss three different models I plan on pursuing in more detail. Currently, I have just hashed out their structure and intuit-ed their predictions. After collecting more data, I plan on implementing them directly.

3.1 Model 1

A very basic model for the above experiment and people's rankings would be a very simple Bayesian model. In this model, I posit that, after viewing the initial freeze frame, there is some probability distribution over what surface it is. Furthermore, for each surface, there is a likelihood distribution over loudness for the surface (which I would model as a Gaussian). Then, we can compute the loudness distribution for a trial using Bayes rule: $P(loudness) = \sum_{surface} P(loudness|surface)P(surface)$. This is actually just a basic Gaussian Mixture Model (anchored on the scene viewed). In terms of a generative process, this could be thought of as some uniform Dirichlet prior on $P(surface)$, perhaps $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, and then the surface generating the scene and sound with certain probabilities. For the data above, this model could be a reasonable fit. However, in general, if there is a difference in incongruent stimuli, this model will fail to capture that trend.

3.2 Model 2

To account for possible differences in the incongruent stimuli, I propose a slightly more complicated model. In this model, we have a generative process as follows. We start by picking a true sound. Then, we generate a probability distribution over possible surfaces that could've generated this sound (note this is equivalent to doing posterior inference in model 1, but we could also just explicitly write these distributions out). Then, we have the observed sound, which is dependent on both the surface and the true sound. One can think of this as a Bayes Net with 3 nodes A, B, C, with A having arrows to B and C, and B having an arrow to C. If we take the mean field assumption for C (namely that $P(C|A, B) = P(C|A)P(C|B)$) and again model the two distributions as Gaussians, we note that the incongruent stimuli would actually yield loudness judgments somewhere between the two signals. For example, if the surface was wood but the sound was ceramic, this model (under the mean field assumption) would dictate that the perceived loudness is somewhere between the distrubtion for wood and the distribution for ceramic. This is observed in subject 8 above. Although this doesn't match our prediction about physics-based audiovisual simulation, if this model does end up fitting the larger dataset better, it could give insight into how humans deal with incongruent stimuli.

3.3 Steps Forward

Finally, the model I am most interested in pursuing, but is also the most complicated, would be an extension of the HumanGalileo project. This extension would incorporate audio stimuli in the belief updating architecture. I plan on pursuing this track as I continue my independent research with Ilker Yildirim.

Code

The code used in this project can be found at: https://github.mit.edu/ilkery/automatic-intuitive-physics. This is currently a private repository, but it will soon be made public (since Ilker created the repository, he controls the settings... I've emailed him about this).

References

[1] Battaglia, P.W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. PNAS, 110(45), 18327–18332.

[2] Shergill, S.S., Bays, P. M., Frith, C.D., & Wolpert, D.M. (2003). Two Eyes for an Eye: The Neuroscience of Force Escalation. Science, 301, 187.

[3] Blakemore, S., Wolpert, D., & Frith, C. (2000). Why can't you tickle yourself? NeuroReport, 11(11), R11-15