Shape and Material from Sound

Zhoutong Zhang, Qiujia Li, Zhengjia Huang, Jiajun Wu, Joshua B. Tenenbaum, and William T. Freeman

NIPS 2017

Chorom Hamm

crhamm@mmlab.snu.ac.kr

Feb 21, 2023

Outline

- Introduction
- System Design and Details
- Performance Evaluation
- Conclusion

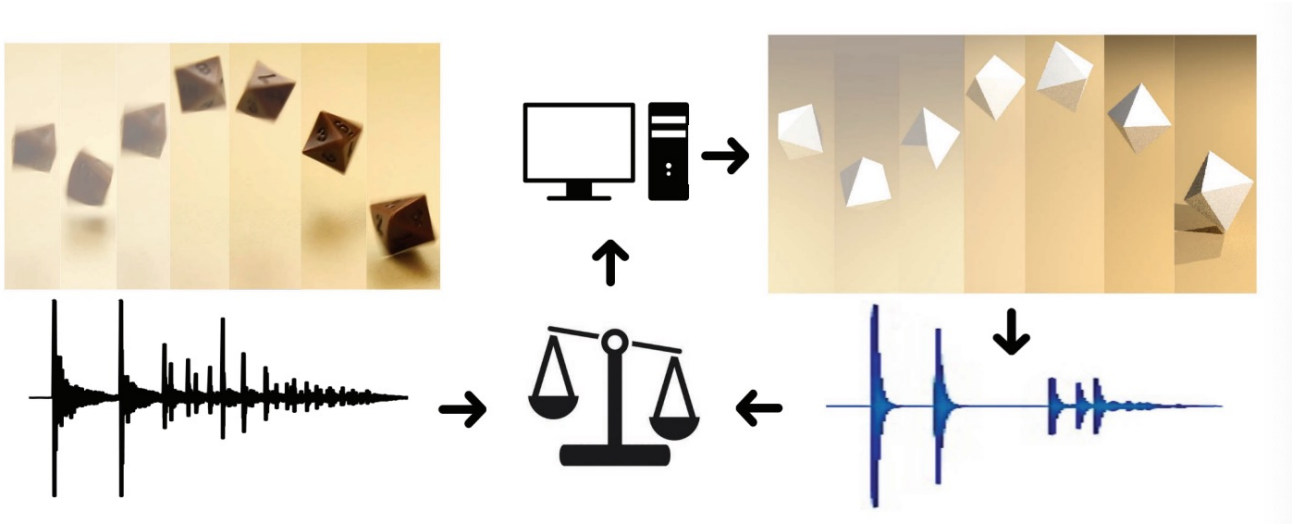
Motivation

- Hearing the sound when interacting objects
  - Human ears can recognize the number of objects involved, materials and surface smoothness… (reference from Zwicker and Fastl, 2013 and Siegel et al., 2014)
- Judging somethings by human
  - Mental physics engine provides probabilistic simulations based on knowledge of physical properties and basic laws of physical interaction *(reference from Battaglia et al., 2013)*

- ***Is it possible to build machines to mimic the above process?***

Introduction

- Building a physics-based simulation engine
- Presenting an analysis-by-synthesis approach for inference

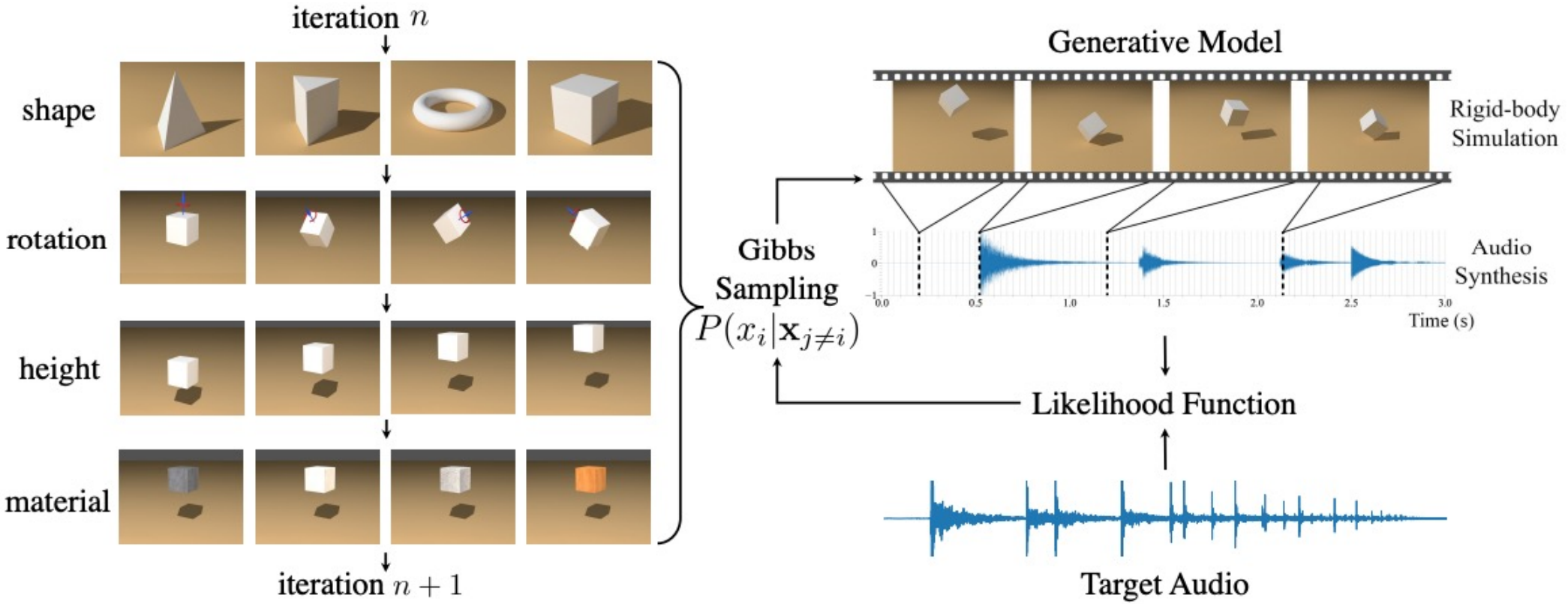


- Accelerating the process by learning a map from sound to object
- Achieving the near-human performance in terms of object shape, material, and falling height

Related Work

- Human visual and auditory perception
  - How humans can infer object properties like shape, material, size from audio
- Sound simulation
  - Simulating object vibration using the finite element method and approximating it
- Physical object perception
  - Understanding physical object properties like mass and friction
- Analysis-by-synthesis
  - Generative models with data-driven proposal focusing on auditory data

System Overview



Audio Synthesis Engine

- Simulating the physical process to generate realistic sound
- Rigid body simulation
  - Approximating the sound based on given an object's 3D position and orientation, mass, and restitution by simulating the interaction between an object and the environment
- Audio synthesis
  - offline module for computing object's intrinsic properties such as shape and Young's modulus for fast synthesis
  - online module to produce the simulated sound by measuring the pressure changes of vibration

Acceleration of Audio Synthesis

- To accelerate the analysis-by-synthesis inference to near real-time
- Select the most significant modes until total energy reaches 90%
- Stop the process based on decreasing the amplitude of the damped sound
- Parallelize the synthesis process to compute each on an independent thread and integrate them according to the timestamps

Settings	Time (s)
Original algorithm	30.4
Amplitude cutoff	24.5
Principal modes	12.7
Multi-threading	1.5
All	**0.8**

Variables of Generative Model

- Constructing an audio dataset with 14 primitives and 10 specific moduli
- After pre-computing, generating synthetic audio clips in a near real-time
- Setting the total simulation time to 3 seconds

Variable	Range	C/T	Variable	Range	C/T
Primitive shape (s)	14 classes	D	Specific modulus (E/ρ)	$[1, 30] \times 10^6$	D
Height (z)	$[1, 2]$	C	Restitution (e)	$[0.6, 0.9]$	C
Rotation axis (i, j, k)	S^2	C	Rotation angle (w)	$[-\pi, \pi)$	C
Rayleigh damping (α)	$10^{[-8,-5]}$	C	Rayleigh damping (β)	$2^{[0,5]}$	C

Inference Models

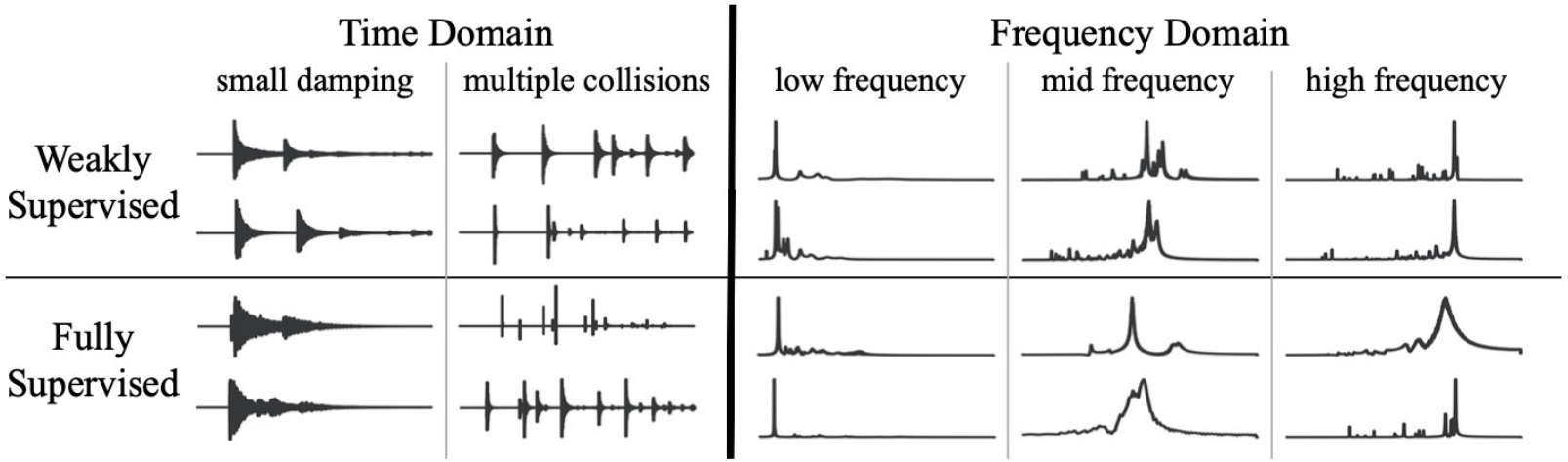
- Unsupervised model as the first inference model to approach it similarly to the way how humans infer
  - Purpose: to recover the latent variables to best reproduce the sound
  - Ways: adopting Gibbs sampling to find x that maximizes the likelihood function
  - Inputs: one test case with no annotation
  - Features: Use the time difference for height and spectrogram for others

Acceleration of Sampling Process

- Self-supervised model as the second inference model using past experiences obtained from unsupervised inference models
  - Purpose: to accelerate the sampling process
  - Ways: Training a deep neural network labels generated by unsupervised inference model and starting the sampling process from a better initialization position
  - Inputs: one test case with labels from the first step
  - Features: same as the first model

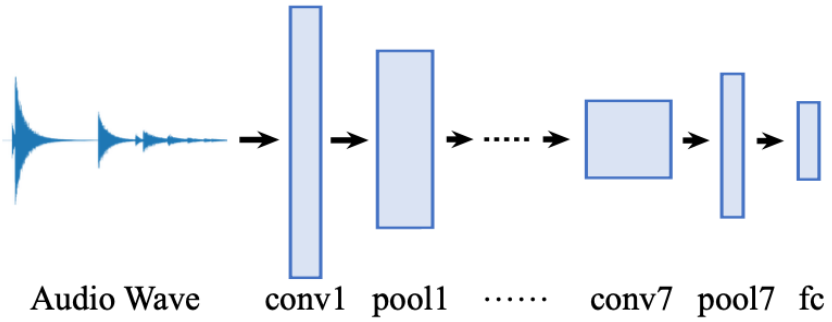
Comparison of Inference Models

- Weakly-supervised learning with coarsen ground truth labels
  - Shape: with edge, with curved surface, and pointy
  - Material: steel, ceramic, polystyrene, and wood
  - Height: low and high



Contrasting Model Performance

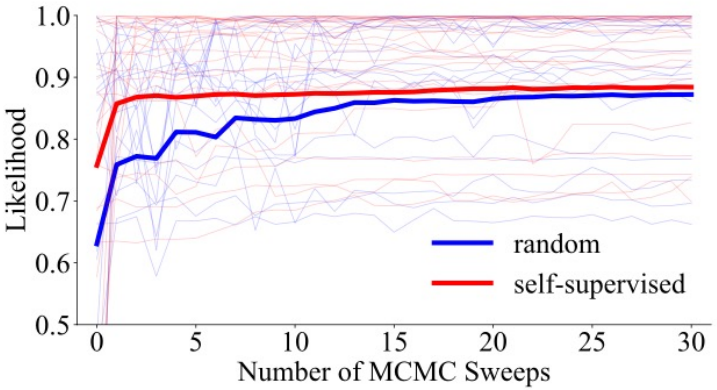
- Purpose: to study how past experiences or coarse labeling can improve the unsupervised results
- Ways
  - Sampling setup: perform 80 sweeps of MCMC sampling over all the latent variables (Discrete: uniform distribution / Continuous: auxiliary Gaussian variable)
  - Deep learning setup: Use architecture of SoundNet-8 to reproduce a 1024-dim feature vector



Inference Result

- Shape and specific modulus: classification accuracy
- Height, Rayleigh damping coefficients, and restitution: normalized MSE

Inference Model		Latent Variables				
		shape	mod.	height	α	β
Unsupervised	initial	8%	10%	0.179	0.144	0.161
	final	54%	56%	0.003	0.069	0.173
Self-supervised	initial	14%	16%	0.060	0.092	0.096
	final	52%	62%	0.005	0.061	0.117
Weakly supervised	initial	18%	12%	0.018	0.077	0.095
	final	62%	66%	0.005	0.055	0.153
Fully supervised	final	98%	100%	0.001	0.001	0.051



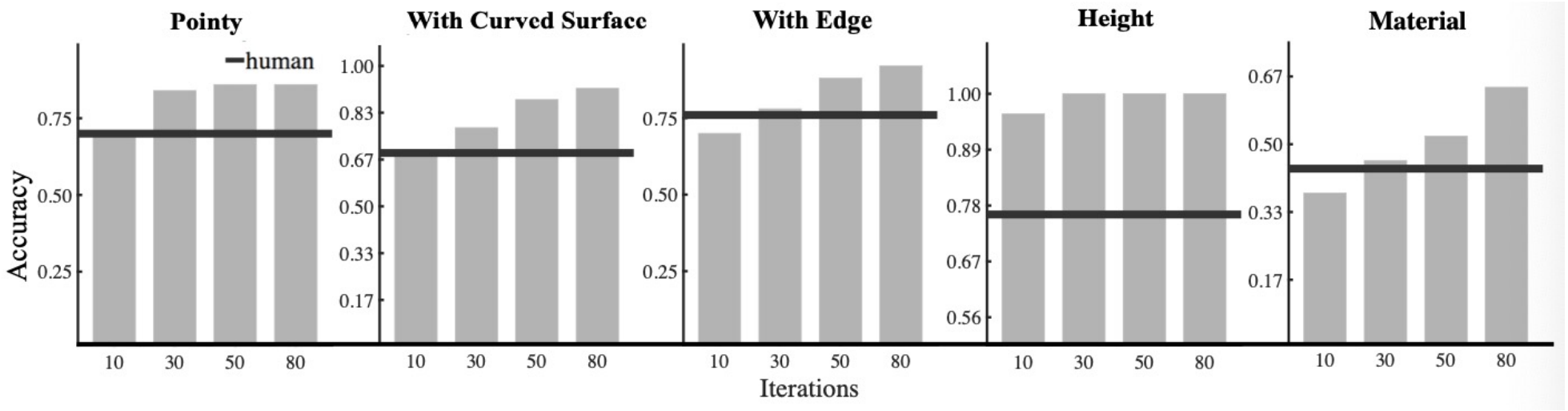
Evaluations

- To evaluate the performance of inference procedure compared to humans' ability
  - Evaluation using synthetic audio with ground truth labels
  - Inference performance with real-world recordings

- Inferring the object's shape, material and height fallen from the sound
  - Classification with the labels used by weakly-supervised model
  - Experiments with randomly selected 52 test cases
  - Response collection of 192 for shape, 556 for material, and 492 for height

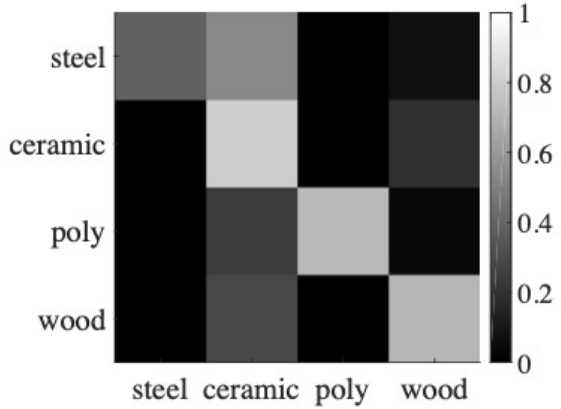
Evaluations – Shapes

- In case of humans, three binary judgments can be distinctive: “with edge”, “with curved surface”, and “pointy”
- In case of inference model, it achieves around the same level when running the unsupervised algorithm for 10-30 iterations

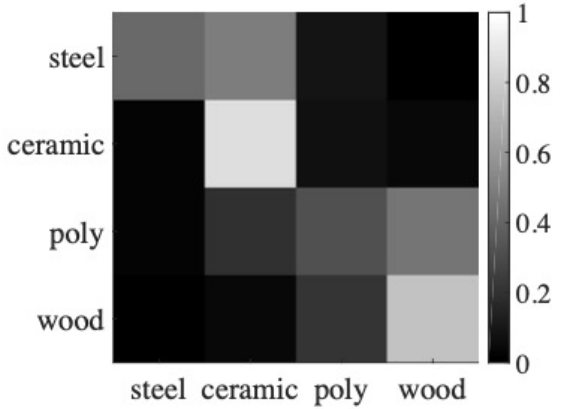


Evaluations – Materials

- To classify four possible materials: steel, ceramic, polystyrene and wood, model needs to distinguish based on density, Young's modulus and damping coefficients
- Hard to classify when having similar damping and specific modulus



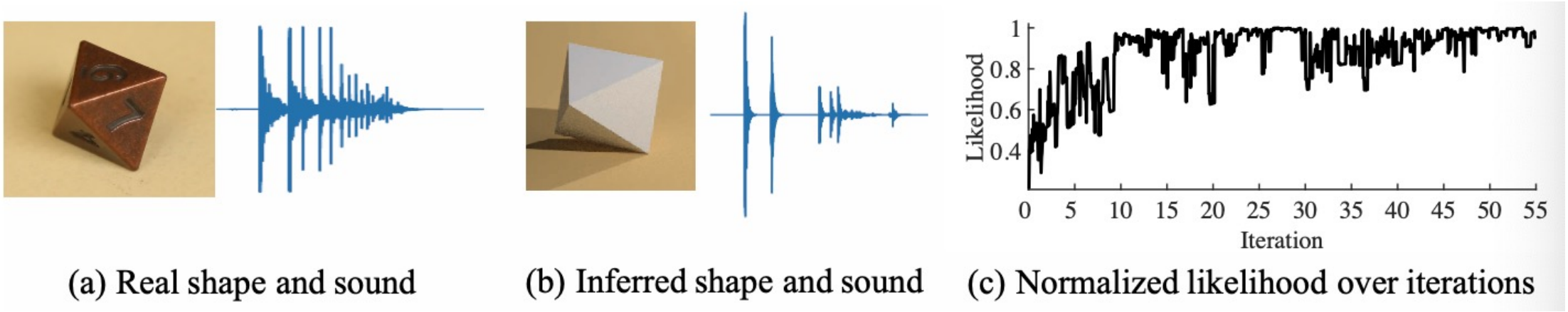
Unsupervised model



Human

Evaluations – Real Scenes

- Classify three candidate shapes: tetrahedron, octahedron and dodecahedron
- Need a more robust feature and a metric like energy distribution with EMD (Earth Mover's Distance) due to highly complex factors in the real-world scenarios



(a) Real shape and sound (b) Inferred shape and sound (c) Normalized likelihood over iterations

Conclusion

- Proposing a model to estimate physical properties of objects from auditory inputs
- Incorporating the feedback of an audio synthesis engine in the loop
- Demonstrating to accelerate inference with fast recognition models
- Comparing model predictions with human responses on judgement tasks

Thank you