A Computational Model for Combinatorial Generalization in Physical Auditory Perception

Yunyun Wang[1,2], Chuang Gan[4], Max H. Siegel[1], Zhoutong Zhang[3], Jiajun Wu[3], Joshua B. Tenenbaum[1,3,5]
{wyy, chuangg, maxs, ztzhang, jiajunwu, jbt}@mit.edu
[1]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, 02139, USA
[2]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, 100084, China
[3]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, 02139, USA
[4]MIT-IBM Watson AI Lab, Cambridge, MA, 02142, USA
[5]MIT Center for Brain, Minds and Machines, Cambridge, MA, 02139, USA

Abstract

Humans possess the unique ability of combinatorial generalization in auditory perception: given novel auditory stimuli, humans perform auditory scene analysis and infer causal physical interactions based on prior knowledge. Could we build a computational model that achieves combinatorial generalization? In this paper, we present a case study on box-shaking: having heard only the sound of a single ball moving in a box, we seek to interpret the sound of two or three balls of different materials. To solve this task, we propose a hybrid model with two components: a neural network for perception, and a physical audio engine for simulation. We use the outcome of the network as an initial guess and perform MCMC sampling with the audio engine to improve the result. Combining neural networks with a physical audio engine, our hybrid model achieves combinatorial generalization efficiently and accurately in auditory scene perception.

Keywords: auditory scene analysis; combinatorial generalization; physical simulation

Introduction

Humans engage with new auditory scenes every day, but we manage to interpret them effortlessly. For example, when we hear a mixture of sounds that we've never heard before—a person talking, a bird singing and a train on tracks—we perceive each component distinctly and clearly. How is this possible? We speculate that humans perceive complex mixtures by composing previously heard sounds together, combinatorially. We refer to this hypothesized ability as *combinatorial generalization*.

In recent years, researchers have developed multiple tools to try to understand the perceptual world in the same way as humans do. In both visual and auditory fields, neural network have shown impressive perceptual abilities. However, neural networks lack the ability of combinatorial generalization: they cannot recognize categories for which they have not been trained. Indeed, a neural network does not contain any concepts in its ontology besides those in the training labels. Therefore, we turn our attention to how humans achieve combinatorial generalization and attempt to build a computational model for auditory scene perception.

Recent behavioral studies (Battaglia, Hamrick, & Tenenbaum, 2013; Sanborn, Mansinghka, & Griffiths, 2013) suggest that intuitive physics (the human perception of physics) may be modeled using a probabilistic physics engine. Following these seminal works, we built a hybrid inference engine which frees us from the aforementioned restrictions of neural networks. The specific sampling algorithm that we use is MCMC sampling with the Metropolis-Hastings acceptance rule. We set up a box-shaking scenario in which a box containing multiple balls, of potentially different materials, are shaking in a box and the materials must be inferred. We generated synthetic audio using an audio synthesis engine (described below); the same engine is also used during sampling.

Our model consists of two parts. First, we use a neural network for direct perception. We train the network using labeled synthetic audio from a scene of shaking a box with only one ball inside. We then use the network to obtain probabilistic prediction of ball materials in scenes with two or three balls. These estimates are used as an initial guess of the materials. Second, we perform MCMC sampling to iteratively update the balls' materials, generate corresponding audio using the physics engine, and compare the generated audio with the observed audio with a perceptual distance—sound texture distance (McDermott & Simoncelli, 2011). A likelihood function is used to decide whether we accept the new material or not. After multiple steps, the outcome becomes stable and achieves a high accuracy. Combining neural nets and simulation-based sampling, our hybrid model achieves combinatorial generalization in physical auditory perception.

Related Work

Human auditory perception The fields of psychoacoustics and auditory perception and cognition have a long history; see (Kunkler-Peck & Turvey, 2000; Zwicker & Fastl, 2013) for reviews.

Sound synthesis James, Barbič, and Pai (2006) proposed a method for modal sound synthesis - approximating the vibrational modes of objects by solving the Helmholtz equation with the Boundary Element Method. Zhang, Wu, et al. (2017) achieved large-scale sound synthesis and accelerated the process to near real-time performance. We modified the framework proposed by Zhang, Li, et al. (2017) and applied controlled motion restriction to the object to apply a real shak-

ing motion.

Analysis-by-synthesis Our method is related to attempts to understand physical perceptual scenes with generative models, in other words, analysis-by-synthesis. Wu, Yildirim, Lim, Freeman, and Tenenbaum (2015); Zhang, Li, et al. (2017) studied inference of latent physics parameters from visual and auditory data respectively. Our work focuses specifically on the combinatorial generalization ability of analysis-by-synthesis approaches, rather than on extracting latent variables from data drawn from the same distribution as the model was trained on.

Setup

Problem Setup

To demonstrate that our hybrid model is capable of doing combinatorial generalization, we designed a box-shaking game. A number of balls with various materials are put inside a wooden box and the box is shaken to produce sound. There are four types of materials (polystyrene, wood, bronze, aluminum) in total and their collision sounds differ. The models aim to learn the materials by training with the audios from the one-ball scenario and generalize to recognize the materials in two or three balls scenarios.

The box is initially placed horizontally and the balls are randomly placed inside the box. We simulate the process using the audio synthesis engine described below and generate the corresponding audio. The training data is a collection of 400 one-ball-shaking scenarios each with a random initialization for the initial position of the ball, introducing a difference in the generated audio. The training data is balanced with 100 cases for each material.

Audio Synthesis Engine

In real life, obtaining clear audio requires a strictly controlled environment and it's almost impossible to keep extraneous variables constant (e.g. the motion that produces the sound). Training neural networks demand large amounts of labeled data, which makes the data collection process time-consuming. Therefore, we chose to use a realistic sound synthesizer to produce the audio for the model and human study to avoid the previous problems. Audio synthesis engine produces realistic sound by simulating the physics evolution of a given system in the following steps.

Controlled rigid body simulation. Collision information is crucial to synthesizing realistic sound. The physics engine Bullet (Coumans, 2015) is able to simulate the motion of the objects in a sequence of time, given the initial position, orientation, velocity of the objects. To successfully imitate the shaking motion, we control the exact motion of the shaking box by setting the position and orientation of the box at an extremely high frequency, a time step of 1/3600 second. The simulation goes at the same rate and the other objects are free to interact according to the engine. The physics engine outputs the location, magnitude and direction of the collisions at each time step for audio synthesis, along with the positions of each object for visualization.

To capture the realistic trajectory of a natural shaking motion, we used OptiTrack V120 motion tracking system to get the box's position in Euclidean coordinates and rotation in quaternions at a frame rate of 120 FPS. The data is then linearly interpolated to a frame rate of 3600 FPS to fit the simulation rate. An up-and-down shaking trajectory is used uniformly throughout the experiments. An illustrative example can be found in Figure 1.

Audio synthesis. We adopt the method employed by previous work on realistic sound synthesis, SoundSynth (Zhang, Li, et al., 2017). For a specific shape with fixed Young's modulus, SoundSynth allows offline pre-calculation on the object. The object's vibrational modes are generated by finite element methods (FEM) and boundary element methods (BEM) is applied to solve the Helmholtz equation conditioned on the vibration modes. After receiving the collision, position, and amplitude from the rigid body simulator, SoundSynth computes the impulses on the tetrahedral meshes of the object and combines them with the pre-calculated vibration modes to produce sound.

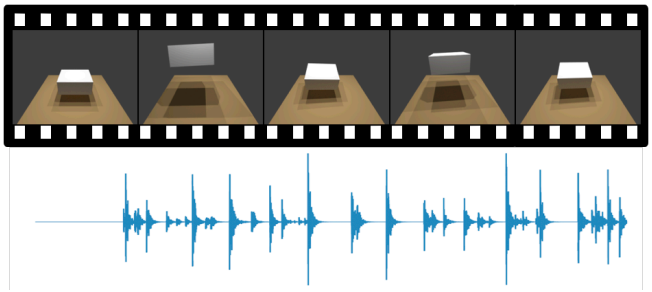

Figure 1: An example of the shaking motion and the generated audio.

Models

Neural Networks

Recognizing one material from sound is often considered as a pattern recognition process in which the model transforms collected perceptual data to statistics and makes future judgments by comparing new data with memorized statistics. This can also be viewed as a classification problem and can be effectively solved by training a neural network. However, whether the neural network is simply fitting the labels or capturing the key information about the data is unclear. Therefore, we adapt the deep learning network VGGish (Hershey et al., 2017) as our baseline to investigate the neural network's performance on the combinatorial generalization problem. The architecture of the network can be found in Figure 2.

The original VGGish network transforms the audio waveform to a spectrogram and outputs a 128-dim feature vector. The feature vector is fully connected with a 100-dim layer and then fully connected to a four-class layer with sigmoid activa-

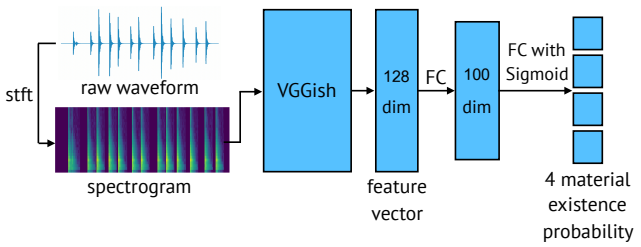

Figure 2: The architecture of our baseline network.

tions. The highest probability materials are selected as the output. We used the pretrained VGGish (Gemmeke et al., 2017) and fine-tuned the network using the Adam optimizer (Kingma & Ba, 2015) with a batch size of 20, a learning rate of 0.0001, and the cross-entropy loss.

Physical Simulation

We would like to imitate how human handle this type of auditory perception. It is hypothesized that human's comprehensive understanding of the auditory scene requires not only direct pattern recognition but also generation of hypotheses and comparison with the heard audio. We model this process as MCMC sampling with simulation and the Metropolis-Hastings acceptance rule. Figure 3 shows the sampling pipeline.

For a scene with N balls, let M denote a material vector of N materials. During each step t, we update M^t through probability distribution $\hat{p}(M_i^t|M_1^t...M_{i-1}^t M_{i+1}^{t-1}...M_N^{t-1})$. Since materials are independent of each other and are uniformly random chosen, the update is equivalent to randomly selecting a new material. We aim to find a M that maximize the likelihood between the audio produced by M which can be simulated with our synthesis engine and the ground truth audio A.

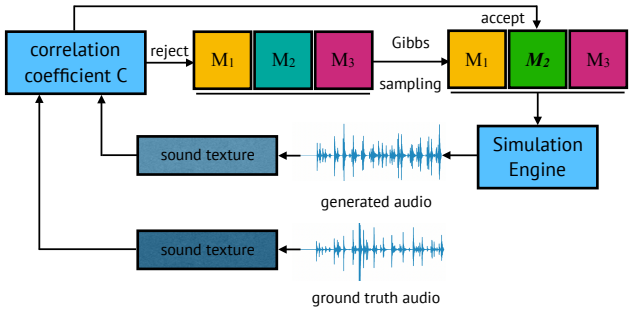

Figure 3: Flowchart of updating one material in one step of MCMC sampling in a three-ball scene.

The likelihood should be measured through a loss function that reflects how humans judge auditory similarity. In this case, we adopt sound texture (McDermott & Simoncelli, 2011), a reasonable model for human sound similarity. The sound texture method decomposes the audio into different frequency bands and calculates statistics of cochlear envelopes. The texture distance is measured by computing the correlation coefficient C between the statistic vectors of the two audios, indicating the similarity of two audios. Thus C does not depend on the collision sequence, mitigating the stochasticity introduced from the random initial positions. We map the correlation coefficient C to an exponential function $\mathcal{L}(C)$ as our likelihood function,

$$\mathcal{L}(C) = e^{(C-1)\cdot(2.5+t/4)}, \tag{1}$$

where t is the current step of MCMC sampling. The likelihood function controls the acceptance of the newly generated materials according to the correlation and time. We follow the Metropolis-Hastings algorithm and each update accept new materials with probability $\min\left(1, \frac{\mathcal{L}(C)}{\mathcal{L}(C_{last})}\right)$, otherwise reject and keep the old materials.

Results

We measure the accuracy of inference of multiple balls by doing maximum matching between two material vectors.

Let $\mathcal{D}$ denote the maximum matching of the material vectors, and we define the accuracy:

$$\mathcal{A}(M, \bar{M}) = \frac{|\mathcal{D}(M, \bar{M})|}{N} \tag{2}$$

The simulation-based sampling model requires an initial material vector. We initialize the material vector with the network model outcome and random label respectively. As sampling is a random process, we reduce the uncertainty of the result by averaging the outcome of 5 sessions and running long enough steps until it converges. The final accuracy and standard error of the hybrid model can be seen in Table 1.

# balls	neural network	Hybrid model
2	0.73	**0.90** $\pm$ 0.044
3	0.67	**0.82** $\pm$ 0.042

Table 1: Average accuracy of the neural networks and our hybrid model. Our hybrid model is initialized with neural network outcome and run for 30 sampling steps.

Comparison between models. We tested the neural network that achieves 100% one-ball test accuracy on the two-ball and three-ball scene but it scores poorly. Our model reaches high accuracy regardless of the initialization. Initialization with network model reaches high accuracy faster because neural network offers basic perception about the data which may be partially accurate. Both initialization methods stay at a high accuracy in the end which indicates the sampling is stable after relatively long steps.

Comparison between scenes. Recognizing materials of three balls is harder than that of two balls because, aside from recognizing the appeared material types, the numerosity also matters. Inferring numerosity may be hard for sound texture as the difference of the sound due to numerosity might not

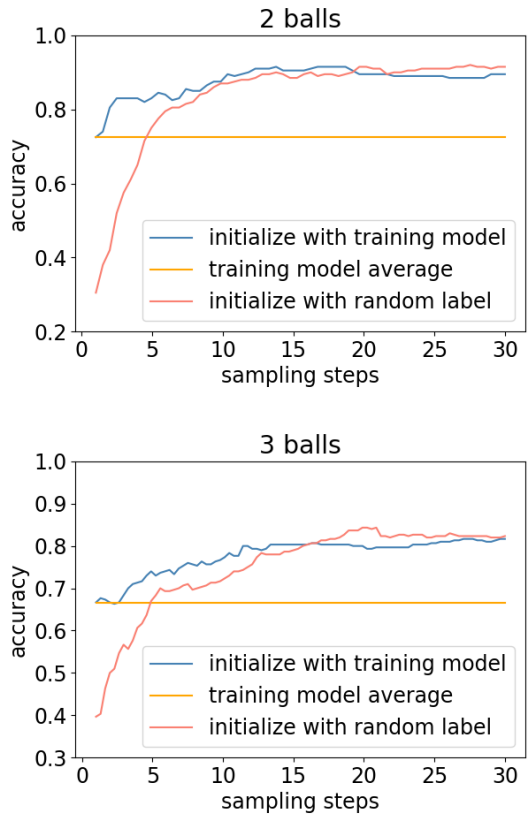

Figure 4: The performance of the MCMC sampling with different initializations.

be that noticeable. This might explain the accuracy drop in the three-ball scene as compared to the two-ball scene. The convergence speed of the two scenes differs as well. Two-balls scene reaches convergence state at around 10 steps while three-balls scene reaches around 20 steps (Figure 4). Three-balls scene requires recognizing all materials correctly to have a high correlation coefficient which takes more steps to achieve. This result indicates that combinatorial generalization becomes more difficult as the number of objects increases.

Overall, our hybrid model combines the neural network and simulation-based sampling, leading to more efficient and more accurate results in combinatorial generalization.

Conclusion

In this paper, we designed a hybrid model and demonstrated its ability to perform combinatorial generalization. The neural network baseline performs poorly on this task, while our method—using the neural network output as initialization for simulation-based sampling—achieves significantly better performance even with relative few sampling steps. The hybrid model puts forward a new way to refine physical perception with a physics engine and to blur the time-sensitive audio to recognizable human-similar texture with sound texture. This idea may be extended to other auditory perception scenarios.

Acknowledgements

This work was supported in part by the Center for Brains, Minds and Machines (CBMM, NSF STC award CCF-1231216), ONR MURI N00014-16-1-2007, IBM, and Facebook. We thank Chi Han and Zhezheng Luo for discussions.

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences (PNAS)*, *110*(45), 18327–18332.

Coumans, E. (2015). Bullet physics simulation. In *ACM SIGGRAPH 2015 Courses.* ACM.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., ... Wilson, K. (2017). Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

James, D. L., Barbič, J., & Pai, D. K. (2006). Precomputed acoustic transfer: output-sensitive, accurate sound generation for geometrically complex vibration sources. In *ACM Transactions on Graphics (TOG)* (Vol. 25, pp. 987–995).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations.*

Kunkler-Peck, A. J., & Turvey, M. (2000). Hearing shape. *Journal of Experimental psychology: Human Perception and Performance*, *26*(1), 279.

McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, *71*(5), 926–940.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems* (pp. 127–135).

Zhang, Z., Li, Q., Huang, Z., Wu, J., Tenenbaum, J. B., & Freeman, W. T. (2017). Shape and material from sound. In *Advances in Neural Information Processing Systems* (pp. 1278–1288).

Zhang, Z., Wu, J., Li, Q., Huang, Z., Traer, J., McDermott, J. H., ... Freeman, W. T. (2017). Generative modeling of audible shapes for object perception. In *The IEEE International Conference on Computer Vision (ICCV).*

Zwicker, E., & Fastl, H. (2013). *Psychoacoustics: Facts and models* (Vol. 22). Springer Science & Business Media.