Sample-efficient learning of auditory object representations using differentiable IR synthesis

Vinayak Agarwal [1 2] **James Traer** [3] **Josh H. McDermott** [2]

Abstract

Many of the sounds we hear in daily life are generated by contact between objects. Rigid objects are often well approximated as linear systems, such that IRs can be used to predict their vibrational behavior. IRs carry information about material and shape. Previous research has shown that IRs measured from objects can be used to generate realistic impact, scraping and rolling sounds. However, it has been unclear how to efficiently synthesize IRs for objects of a particular material and size. Here we present an analysis-by-synthesis technique that uses a differentiable IR synthesis model to infer generative parameters of a measured IR. Then, we introduce a way of representing auditory material as distributions in the generative parameter space. Object IRs can be sampled from these distributions to render convincingly realistic contact sounds.

1. Introduction

Imagine you hear the sound of something falling on the floor in the next room. What was it? Might it have damaged the floor, or hurt someone it hit on the way down? Such physical interactions between objects are important for humans and machines to perceive correctly. Interacting objects produce sounds which encode information about physical variables - the motion of the objects, their shapes, and their materials (Gaver, 1993a;b). Yet we know relatively little about how humans derive information about physical interactions from sound, and lack machine systems to mirror our abilities (Bianco et al., 2019).

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, US [2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, US [3]Department of Psychology, University of Iowa, Iowa city, USA. Correspondence to: Vinayak Agarwal <vinayaka@mit.edu>.

Published at the Differentiable Almost Everything Workshop of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. July 2023.

Recent progress in contact sound synthesis has made it possible to realistically render the sounds arising from physical interactions such as impacts, scrapes and rolls (Rocchesso et al., 2003; Rath & Rocchesso, 2005; Aramaki & Kronland-Martinet, 2006; Agarwal et al., 2021). Most such sounds can be modelled using linear systems that represent vibrating objects using their IRs, which can be measured from individual objects (van de Doel & Pai, 1996). The IRs are excited by forces that occur between objects when they interact, yielding sound (Sinha, 1992).

Although sounds can be synthesized from measured IRs, we have thus far lacked an efficient way to synthesize IRs to obtain new examples of a given material. We sought to estimate distributions over object IRs that could be used for synthesis, as well as for material classification, and potentially as models of human material perception (Klatzky et al., 2000). Because this is a domain where we lack extensive sets of data, it seemed important to represent IRs with a structured model that would be governed by a modest set of parameters, and to be able to infer these parameters from measured IRs (Avanzini & Rocchesso, 2001). Moreover, a differentiable forward model can enable quick and robust inference akin to similar successes in other domains (Hu et al., 2020; Murthy et al., 2021; Clarke et al., 2021).

In this paper, we present a novel differentiable synthesis model for object IRs that combines two signals corresponding to the sinusoidal resonant modes and the stochastic part of the signal. We also present an inference scheme which when used in conjunction with the differentiable synthesis simultaneously infers the stochastic and sinusoidal parts of recorded IRs. Lastly, we present a learned distributions of auditory materials using this inferred space of generative parameters.

2. Methods

2.1. Differentiable IR Synthesis

Building on prior work, we model object IRs as a sum of decaying sinusoidal modes and a noisy transient (Aramaki & Kronland-Martinet, 2006; Ren et al., 2013). The key assumption is that the functional form of the decay is exponential and hence can be modelled with two parameters per

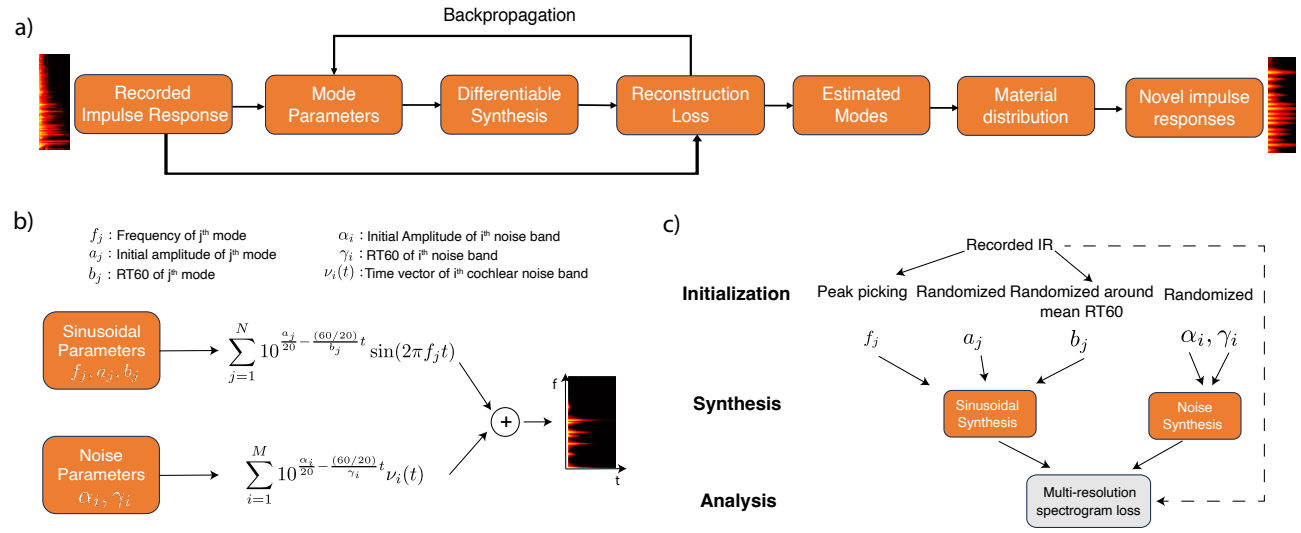

Figure 1. a) Overall scheme for generating novel IRs for a material class. b) Synthesis of IRs from generative parameters in the proposed synthesis algorithm. c) Analysis-by-synthesis scheme for inferring the generative parameters of an IR.

mode: an initial amplitude and a decay rate.

$$y_{ir}(t) = y_{sin}(t) + y_{noise}(t) \tag{1}$$

where $y_i r(t)$ is the net waveform of an IR and $y_{sin}(t)$ and $y_{noise}(t)$ are the decaying sinusoidal modes and the noise transient, respectively.

The decaying sinusoidal modes can be expressed as:

$$y_{sin}(t) = \sum_{j=1}^{N} 10^{\frac{a_j}{20} - \frac{(60/20)}{b_j} t} \sin(2\pi f_j t) \tag{2}$$

where a_j is the initial amplitude (dB), b_j is the RT60 (sec) and f_j is the frequency (Hz) of the j^{th} mode respectively. For the purpose of this work, we only considered the 10 most salient sinusoidal modes for each IR recording ($N = 10$).

The noise transient is modelled as a sum of decaying noise bands:

$$y_{noise}(t) = \sum_{i=1}^{M} 10^{\frac{\alpha_i}{20} - \frac{(60/20)}{\gamma_i} t} \nu_i(t) \tag{3}$$

where α_i denotes the initial amplitude (in dB re: a maximum possible value) and γ_i denotes the RT60 (in sec) for the amplitude envelope on the i^{th} cochlear noise band. We used $M = 10$ noise bands to model the transient, as this seemed sufficient to account for the variations in spectral shape in everyday IRs. The noise bands were generated by filtering Gaussian white noise by a simulated cochlear filter bank, with the intention of making the spectral detail equally discriminable to humans across the spectrum. Specifically, we first generated a random Gaussian noise sample and then filtered it using FIR filters whose cutoffs were equally spaced on an ERB scale (Glasberg & Moore, 1990).

We found the noise transient to be critical to modelling the vibrational response of damped, soft materials like plastics, cardboards etc. This component was also important for large objects, for which multiple vibration modes can occur in close frequency proximity, giving rise to a noise-like band in the IR.

2.2. Inference algorithm

Learning low-dimensional object representations has been the central focus of perception research. Having modelled the acoustical regularities of object IRs through a generative model, we wanted to test if recorded object IRs can be expressed in terms of the proposed parameters. If we are able to infer a unique set of parameters within this generative model that can yield compelling resynthesis, then the proposed parameter set can be a useful stimulus-computable representation of object IRs.

Based on the differentiable model described above, we propose the use of gradient-based optimization to achieve this goal. For this, we used a spectrogram-based loss function that quantified the difference between the measured IR and a synthetic replica generated from the model. To infer the generative parameters through gradient-based optimization, we implemented the algorithm shown in Fig. (1) in PyTorch 2.0 using differentiable audio tools contained in TorchAudio toolbox (Paszke et al., 2019).

2.2.1. Initialization

We found that the success of the inference was greatly aided by "good" initial guesses for each parameter. Since the parameters have physical meaning and known regularities, we used physically motivated heuristics to initialize inference.

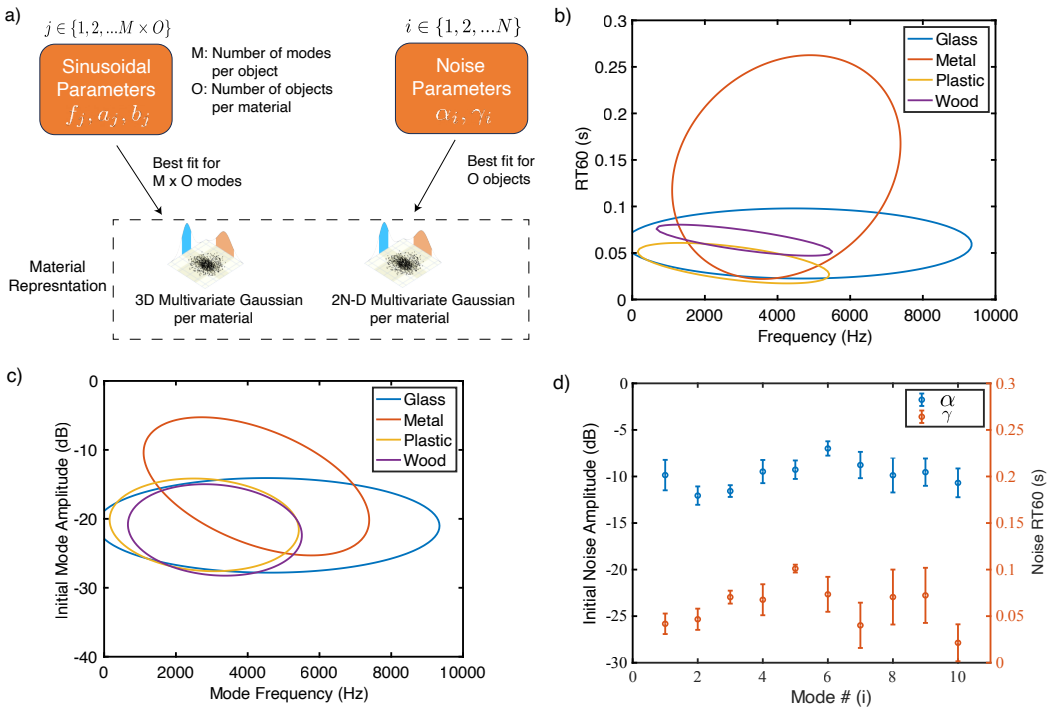

Figure 2. a) Representation of material as two joint distributions (of the sinusoidal and noise generative parameters, respectively). b) Shows the 0.5 probability contours on the RT60 vs frequency marginal distribution on the multivariate Gaussian distribution for different materials c) Shows the 0.5 probability contours for the Initial Amplitude vs frequency marginal distribution of the multivariate normal distribution for different materials d) Example of the noise parameter distribution for wood IRs (bars show standard deviation).

The sinusoidal mode frequencies [f_i, Eq. (2)] were most important to initialize well because when hypothesized mode frequencies are sufficiently different from the true values, small changes to the mode frequencies leave the spectrogram-based reconstruction loss unchanged, preventing gradient-based optimization. We used the power spectrum of the IR to detect possible modes, using peak picking to identify the top N candidate modes with the highest average power. We applied A-weighting to the spectrum prior to peak picking to emphasize perceptually relevant regions of the spectrum (Lee, 1979). To avoid assigning multiple peaks to the same mode, we constrained the selected peaks to be at least 100 Hz apart. To avoid explaining parts of the noise transient with sinusoidal components, we required the peak prominence ≥ 2 (to reject noise bands, that were typically wider than sinusoidal modes).

We initialized sinusoidal mode amplitudes (a_i) by uniformly sampling from [-10 dB, -30 dB]. To initialize mode RT60s (b_i), we first calculate the 'broadband' RT60 for the IR waveform by calculating the time it takes for the average power to dip by 60 dB, and then randomly initialize mode RT60s around the broadband RT60. We found that this helped to make sure that the modes were fit correctly for both resonant and damped materials, and sinusoidal modes were not accounted for by noise bands during optimization.

For the noise parameters, initial guesses for α_j [Eq. (3)] were sampled from a uniform distribution from [-5 dB, -15 dB], while the noise RT60 value (γ_j) were initially sampled from a uniform distribution from [0.04 s, 0.12 s].

2.2.2. Reconstruction Loss

Due to the tradeoff between temporal and spectral resolution in spectrograms, selecting the optimal time window for a spectrogram representation is often a challenge. Larger time windows yield better estimates of sinusoidal mode frequencies where as shorter time windows are better for estimating modal and noise RT60s

To circumvent this issue, we used a multi-resolution spectrogram loss. We calculated four different spectrograms (Number of FFT points - 4096, 1024, 256, 64) for the synthesized and recorded IRs; the loss was the sum of the Huber loss between these representations of the two signals (Hastie et al., 2001).

2.2.3. Optimization scheme

For gradient-based optimization during inference, we used the Adam optimizer in PyTorch 2.0 (Paszke et al., 2019). We used a learning rate of 2×10^{-6}. Larger learning rates produced poor results, presumably because of the multi-dimensional spiky loss landscape.

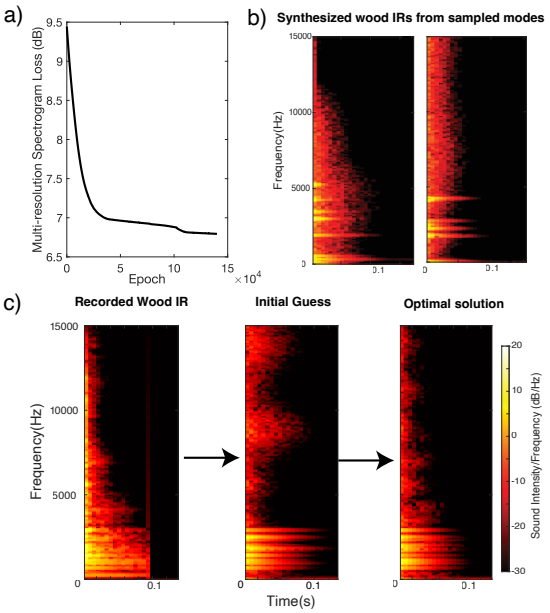

Figure 3. a) Multi-resolution spectrogram loss during an example optimization run. Final loss after 1.4e5 steps was 6.83 dB b) Spectrograms of two example wood IRs synthesized from mode statistics sampled from the wood material distribution learnt using 5 recorded IRs c) Spectrograms of a measured IR (left), the synthetic generated with initial guesses of IR properties (center), and after optimization (right).

2.3. Material representations

The differentiable generative model and inference algorithm allowed us to infer the parameters of a large set of measured IRs. We sought to fit distributions to these inferred parameters in order to classify novel IRs as belonging to one material or another, and to sample novel IRs corresponding to a particular material [Fig. (2)].

We found that multivariate Gaussian distributions were able to capture much of the structure of object IRs. We learned separate distributions for the sinusoidal and the noise parameters. For the sinusoidal parameters $[a_j, b_j, f_j$, Eq. (2)], we fit a 3-dimensional multivariate Gaussian to the mode parameter estimates, pooling across IRs from a particular material class. For the noise parameters, we fit a $2N$-dimensional multivariate Gaussian distribution combining the α and γ parameters for each noise band [Eq. (3)]. Unlike the sinusoidal modes, each noise band has a fixed frequency, and occupied a separate set of two dimensions in the generative space. To generate a sampled IR, we sampled $N = 10$ modes from the sinusoidal distribution, and the $2N$ noise parameter vector from the noise distribution, and then synthesized the IR using the generative model described above.

3. Results and Discussion

3.1. Inference of modes

When applied to a measured IR, the inference scheme typically converged to a low reconstruction loss and was able to infer parameters that resynthesized an IR that was perceptually similar to the measured IR [Fig. (3)].

We found that the covariance between the various generative parameters was important for the realism of IRs. For instance, physics predicts that higher frequency modes of vibration will have shorter RT60s, and the learned material distributions captured this regularity. If the covariance was replaced with a diagonal matrix, eliminating this regularity, the realism was reduced (based on our subjective impressions).

3.2. Sample-efficient learning of auditory material

From our tests, we found that a modest number of recorded IRs were enough to learn a material representation that could generate new IRs that evoke the perception of the same material type. The number of IR recordings could be as low as four to five.

Using the physical regularities captured by the acoustic generative model, we were able to lower the number of generative parameters by several orders of magnitudes compared to other audio representation learning systems (Engel et al., 2020). These structural assumptions which helped us avoid over-fitting to the small set of IR recordings but at the same time, allowed the model to represent the key physical features that acoustically and perceptually define a material (Traer et al., 2019).

4. Future Work

IRs can be used to generate impact sounds, as well as scraping and rolling sounds – the latter two types of sound simply have more complicated excitation forces that are a function of surface textures (Agarwal et al., 2021). Since the proposed generative model of IRs is differentiable, it can be used as a building block in a more complex generative model for contact sounds. In conjunction with recent progress in the synthesis of impacts, scrapes and rolls, the model presented here will enable the development of computational inference algorithms that could be useful in inferring physical interactions from sound, and in explaining the human perception of these interactions.

Acknowledgements

We extend our thanks to the anonymous reviewers for their constructive feedback. We also acknowledge and thank the K Lisa Yang ICON Fellowship and National Science

Foundation for their continued financial support for this research project.

References

Agarwal, V., Cusimano, M., Traer, J., and McDermott, J. Object-based synthesis of scraping and rolling sounds based on non-linear physical constraints. In *2021 24th International Conference on Digital Audio Effects (DAFx)*, pp. 136–143. IEEE, 2021.

Aramaki, M. and Kronland-Martinet, R. Analysis-synthesis of impact sounds by real-time dynamic filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):695–705, 2006.

Avanzini, F. and Rocchesso, D. Controlling material properties in physical models of sounding objects. In *ICMC*, 2001.

Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., and Deledalle, C.-A. Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America*, 146(5):3590–3628, 2019.

Clarke, S., Heravi, N., Rau, M., Gao, R., Wu, J., James, D., and Bohg, J. Diffimpact: Differentiable rendering and identification of impact sounds. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=wVIqlSqKu2D.

Engel, J., Hantrakul, L., Gu, C., and Roberts, A. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*, 2020.

Gaver, W. W. How do we hear in the world? explorations in ecological acoustics. *Ecological psychology*, 5(4):285–313, 1993a.

Gaver, W. W. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29, 1993b.

Glasberg, B. R. and Moore, B. C. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1-2):103–138, 1990.

Hastie, T., Tibshirani, R., and Friedman, J. The elements of statistical learning. springer series in statistics. *New York, NY, USA*, 2001.

Hu, Y., Anderson, L., Li, T.-M., Sun, Q., Carr, N., Ragan-Kelley, J., and Durand, F. Difftaichi: Differentiable programming for physical simulation. In *International Conference on Learning Representations*, 2020.

Klatzky, R. L., Pai, D. K., and Krotkov, E. P. Perception of material from contact sounds. *Presence: Teleoperators & Virtual Environments*, 9(4):399–410, 2000.

Lee, J. B. Band-limited unweighted measurements as first descriptors of noise. *The Journal of the Acoustical Society of America*, 65(6):1583–1584, 1979.

Murthy, J. K., Macklin, M., Golemo, F., Voleti, V., Petrini, L., Weiss, M., Considine, B., Parent-Lévesque, J., Xie, K., Erleben, K., Paull, L., Shkurti, F., Nowrouzezahrai, D., and Fidler, S. gradsim: Differentiable simulation for system identification and visuomotor control. In *International Conference on Learning Representations*, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Rath, M. and Rocchesso, D. Continuous sonic feedback from a rolling ball. Universita degli Studi di Verona, IEEE Computer Society, 2005.

Ren, Z., Yeh, H., and Lin, M. C. Example-guided physically based modal sound synthesis. *ACM Transactions on Graphics (TOG)*, 32(1):1–16, 2013.

Rocchesso, D., Bresin, R., and Fernstrom, M. Sounding objects. *IEEE MultiMedia*, 10(2):42–52, 2003.

Sinha, D. N. Acoustic resonance spectroscopy (ars). *IEEE Potentials*, 11(2):10–13, 1992.

Traer, J., Cusimano, M., and McDermott, J. H. A perceptually inspired generative model of rigid-body contact sounds. *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx-19)*, Sep 2019.

van de Doel, K. and Pai, D. K. Synthesis of shape dependent sounds with physical modeling. Georgia Institute of Technology, 1996.

Supplementary information

Please find the training data and example re-synthesized sounds on the project webpage here - https://mcdermottlab.mit.edu/ICML2023/sound_website.html