RIGID-BODY SOUND SYNTHESIS WITH DIFFERENTIABLE MODAL RESONATORS

Rodrigo Diaz, Ben Hayes, Charalampos Saitis, György Fazekas, Mark Sandler

Centre for Digital Music, Queen Mary University of London

ABSTRACT

Physical models of rigid bodies are used for sound synthesis in applications from virtual environments to music production. Traditional methods, such as modal synthesis, often rely on computationally expensive numerical solvers, while recent deep learning approaches are limited by post-processing of their results. In this work, we present a novel end-to-end framework for training a deep neural network to generate modal resonators for a given 2D shape and material using a bank of differentiable IIR filters. We demonstrate our method on a dataset of synthetic objects but train our model using an audio-domain objective, paving the way for physically-informed synthesers to be learned directly from recordings of real-world objects.

Index Terms— differentiable signal processing, machine learning, sound synthesis, physical modelling

1. INTRODUCTION

The synthesis of contact sounds from rigid bodies and materials has been of continual interest in applications from music production, through sound design, to the rendering of object sounds in virtual environments [1, 2]. For this reason, the problem has been extensively studied, and physically-based numerical methods are often selected for its solution. However, such approaches typically incur a significant computational cost and storage cost for cached solutions if real-time interaction is required. These limitations inhibit flexibility in such applications, where adapting to new object shapes and materials requires prohibitively time-consuming computation of the new solution.

Of these numerical approaches, which ultimately rely on solving the wave equation, the finite difference time domain and the finite element method (FEM) are most commonly used due to their adaptability. FEM solvers are used to precompute the vibrational modes of arbitrarily shaped rigid objects, allowing contact sounds for these objects to be synthesized using modal synthesis.

The computation of the modes is usually posed as a generalized eigenvalue problem using the mass and stiffness matrices of an object. Using the solution, sound can be rendered with an oscillator bank by projecting an impulse onto the modes at each discrete location within the object. This approach to computing an object’s modes can, however, be computationally expensive. Moreover, every time the shape or material characteristics of the object change, the system must be solved again. Thus, numerous approaches have been proposed to accelerate this process [3, 4, 5].

1.1. Data-driven methods

A recent generative method, proposed by Traer et al. [6], used a perceptually derived statistical model to approximate object impulse responses by approximating the modes within some degree of perceptual tolerance. While this method circumvents the need for a numerical solver and uses filters to account for different object interactions, it relies on abstract intermediate representations of an object’s physical characteristics and does not take into account object geometry.

Jin et al. [7, 8] trained a deep neural network to produce an object’s modes, circumventing the need for a FEM solver. Their model predicted eigenvalues and eigenvectors from sparsely voxelized objects, using as supervision the corresponding modes obtained from the solver. Additionally, in order to generalize to different sizes and materials, the authors suggest the adaptive scaling of the modes based on material parameters (except for Poisson’s ratio) and the size of the object. Similarly, our method uses a discretized shape representation as input to a neural network. However, unlike Jin et al.’s model, no post-processing is required to account for different material parameters. Further, our model is not limited to querying modes at discrete positions in the shape, allowing for arbitrary co-ordinate input.

1.2. Differentiable resonators

While it is possible to synthesise modes using an oscillator bank where each damped sinusoid corresponds to an impulsive mode, resonant filter banks are commonly used for synthesis due to their flexibility, as they can be excited with a variety of signals [9]. Realising such a filter bank, however, requires the corresponding eigenvalue and eigenvector problems to be solved in order to derive the filter coefficients. In this work, we propose an end-to-end learning method for generating resonant filter banks without explicitly solving the system.

Our method relies on differentiable infinite impulse response (IIR) filters, which allow a neural network to directly

Acknowledgements

Rodrigo Diaz and Ben Hayes are supported by UK Research and Innovation [grant number EP/S022694/1]. This research utilised Queen Mary’s Apocrita HPC facility, supported by QMUL Research-IT. http://doi.org/10.5281/zenodo.438045
2. METHOD

Fig. 1 provides an overview of our method, illustrating how we sample synthetic data, prepare inputs, and train our model.

2.1. Learnable modules

To allow efficient synthesis of contact sounds at multiple object positions for a given shape, our system contains two separate neural networks which are jointly optimised. The first is a convolutional neural network (CNN), using the EfficientNet-B0 architecture [14], which generates an embedding from two-dimensional shapes. The network takes as input a 2D occupancy grid describing the shape of an object in discretized space and outputs an embedding of 1000 dimensions. For each distinct shape, only a single forward pass through this module is required.

The second learnable module is a multi-layer perceptron (MLP) which takes as input the shape embedding concatenated with material parameters and the coordinates of the vertex at which the shape will be excited. The cached output of the CNN can be reused to evaluate different excitation positions and materials for the same shape, eschewing the need to recompute a shape embedding. The material parameters $\phi = (\rho, E, \nu, \alpha, \beta)$ consist of mass density ($\rho$), Young’s modulus ($E$), Poisson’s ratio ($\nu$), and Rayleigh damping coefficients ($\alpha, \beta$). Both excitation co-ordinates $x$ and material parameters $\phi$ are normalized in the range $[0, 1]$. The MLP’s output is used to parameterize a differentiable IIR filter bank.

2.2. Differentiable filter bank

Each biquad in our differentiable filter bank is parameterized by a single pole $p_{l,m}$ and zero $q_{l,m}$, where $l$ and $m$ correspond to the filter’s index in the parallel and cascade dimensions, respectively. We constrain the remaining pole and zero of each filter to the complex conjugates of these parameters. Each biquad also has a gain parameter $k_{l,m}$ applied to its numerator. In order to encourage filters with strong resonant peaks, we specify a fixed bias for pole and zero parameters, initialising them near the edge and centre, respectively, of the unit circle.

To ensure filter stability, we constrain the complex poles to be within the unit circle following the method described in [15]. Specifically, we apply an activation function $h : \mathbb{C} \rightarrow \mathbb{C}$ defined as:

$$h(p) = \begin{cases} 1 & |p| < 1 \\ 0 & |p| \geq 1 \end{cases}$$
\[ h(p) = \frac{\tanh |p|}{|p|} \cdot p \]

We do not apply such a constraint to the zeros as we are unconcerned with preserving minimum phase. The transfer function of each filter is thus given by:

\[ H_{l,m}(z) = \frac{1 - 2 \Re (q_{l,m}) z^{-1} + |q_{l,m}|^2 z^{-2}}{1 - 2 \Re (h(p_{l,m})) z^{-1} + |h(p_{l,m})|^2 z^{-2}} \]

While training the model, we do not evaluate the filter’s exact recursive implementation, and instead opt, as in recent work on differentiable IIR modelling [12, 13], to approximate the filter response by sampling it in the frequency domain. That is, we evaluate the filter’s transfer function at discrete points \(e^{j2\pi k/N}\) on the unit circle, where \(N\) is the signal length and \(k \in \{0, \ldots, N/2 + 1\}\). Whilst this method incurs a loss in precision, it allows the filter computation to be performed in parallel on a GPU, whilst also sidestepping the aforementioned issues with training stability and exploding gradients.

Our filter bank consists of a flexible number of parallel filters, each consisting of a cascade of some number of second order (biquadratic) sections. The cascade depth \(M\) and number of parallel filters \(L\) are thus tunable hyperparameters of our method, with the overall transfer function of the filter bank given by:

\[ H(z) = \sum_{l=1}^{L} \prod_{m=1}^{M} k_{l,m} H_{l,m}(z) \]  

### 2.3. Loss

We train our model to minimize the following loss function:

\[ L = \lambda \|X_{\text{mel}} - H_{\text{mel}}\|_2^2 + \gamma \| \log X_{\text{mel}} - \log H_{\text{mel}}\|_2^2, \]

where \(\lambda\) and \(\gamma\) are loss scaling hyperparameters, and \(X_{\text{mel}}\) and \(H_{\text{mel}}\) are the mel-scaled magnitude spectra of the discrete Fourier transform (DFT) and filter frequency response, respectively. In our experiments, we used \(\lambda = 1.0,\) and \(\gamma = 0.1.\)

### 2.4. Dataset

To train our model, we generated a synthetic dataset consisting of 500 convex shapes [16] along with their material parameters and audio synthesised using modal synthesis.

In particular, each convex shape was represented by a point set \(P\), which was triangulated using Delaunay triangulation. Subsequently, each triangular surface mesh \(M = (V,E,F)\) with a set of \(V\) vertex positions, \(E\) edges and \(F\) faces was rasterised into a \(64 \times 64\) occupancy grid. Materials were homogeneous and isotropic, and their parameters were sampled uniformly in the intervals \(\rho \in [500, 15000], E \in [8 \times 10^9, 5 \times 10^{10}], \nu \in [0.1, 0.4], \alpha \in [1, 10] \text{ and } \beta \in [3 \times 10^{-7}, 2 \times 10^{-6}.\)

For each shape, 500 random materials were sampled.

### 2.5. Training

We trained our models using the Adam optimiser with an initial learning rate of \(3 \times 10^{-5}\) and an exponential decay of 0.9 every 300 steps. Models were trained until validation loss did not improve for 20k steps. All models were trained with a batch size of 64 on a single NVIDIA RTX A5000 GPU.

### 3. RESULTS

In Fig. 2 we illustrate our network’s predictions for samples from a test dataset of 100 materials and shapes. We observe that our method closely matches the frequency distribution of the target signals while generalising well across shapes. We provide comprehensive examples of audio rendered with our system in the online supplement at http://rodrigodzf.github.io/iir-modal/results.html, including audio from the objects in Fig. 2, and interpolation of shape and material parameters.

Our model appears to produce shorter decay times than those in the target signal for modes of higher frequency. Further, we note that the complex activation in Eqn 1 causes the gradient magnitude to decrease as a pole approaches the unit circle. It is thus possible that for higher frequency modes with
To ascertain our method’s computational performance, we compared its processing time to that of a conventional FEM solver (as implemented in `scikit-fem`). Table 1 lists the processing times of both methods for 96, 426, and 1792 vertices averaged across 100 meshes of different resolutions.

Our method produces filter bank parameters for a set of 96 arbitrary co-ordinates in a mesh in 16.75 milliseconds using a GPU. Subsequently rendering the audio is possible at negligible computational cost using a recursive implementation of the filter bank. In comparison, modal synthesis using the FEM solver takes approximately 83.66 milliseconds to calculate the mode gains and frequencies for a mesh with the same number of vertex positions. The solutions are only available at the discrete coordinate positions of the vertices. Increasing the number of mesh vertices, and thus obtaining solutions at finer steps, increases the duration of processing.

To produce results for a mesh with 1792 vertices, the FEM solver takes slightly over 1.7 seconds, while our method requires only 0.08 seconds. Note that in these experiments we generate a shape embedding even though the occupancy grid does not necessarily change in the model input, and thus this can be considered a worst case scenario bound on the actual runtime of our model. If we cache the shape embedding, the total time is reduced to 0.01 seconds for this number of vertices. Moreover, as our method does not need a re-discretization of the shape, we can query filter parameters at any continuous valued co-ordinate in the shape with the same constant processing time. Further, our method is not restricted to simultaneously querying multiple points in the mesh to render the sound for a single position. Our approach is therefore preferable in real-time and interactive synthesis for arbitrary shapes.

### 3.2. Filter-bank configurations and baseline comparison

To ascertain the optimal tradeoff between the cascade depth $M$ and the number of parallel filters $L$, we compared three filter bank configurations such that the total number of poles was constant, fixed at $M \cdot L = 64$. We evaluated models trained using these configurations on the test dataset using both the frequency domain mean-absolute and mean-squared error. In general, we observed that filter banks with a lower cascade depth and, thus, a lower filter order converged faster.

Table 2 lists the results for these experiments. Convergence was very poor for the highest order filter bank, with dramatically higher errors than either of the other configurations. The 4th order filter bank achieved the best performance across both metrics, suggesting a hybrid cascade/parallel topology is optimal.

Additionally, we present a comparison between our model and a modified version of Deep-Modal [7], which has been adapted to work with 2D grids. The baseline scores on our evaluation metrics are lower mainly because modal frequencies are computed exactly from a subset of material parameters, whereas our method learns to generate resonant poles at the appropriate frequencies from material conditioning.

### 4. CONCLUSION

This paper presented a novel approach for synthesising contact sounds for rigid surfaces of arbitrary shape and material composition. Our method generates parameters for a resonant IIR filter bank from shape, material, and excitation position inputs. The generated filter bank is suitable for real-time implementation, allowing realistic contact sounds to be produced interactively using a variety of different excitation signals.

In this work, we limited shapes to two dimensions, but we believe that extending the proposed method to 3D objects simply requires adapting the shape encoder to use 3D convolutional layers. Additionally, a future adaption of our model could apply a discretization agnostic network [18] to learn from different surface representations. The rest of the pipeline would not require any alteration.

Finally, since the predicted filter bank can be excited with any signal, we intend to explore the possibilities of complex interaction on the object’s surface, along with a qualitative evaluation.
5. REFERENCES


