CHAPTER 1: Audio-Material Reconstruction for Virtualized Reality Using a Probabilistic Damping Model[1]

1.1 Introduction

Modal sound synthesis improves a user's immersion, but it requires accurate real-world material parameters. Damping, which determines the rate at which vibrations and sound decay over time, is crucial in differentiating between different materials. Some parameters, e.g. density and Young's modulus, can be looked up for known materials, but damping behavior can be difficult to identify and parameterize.

Traditionally, damping parameters are selected through laborious human hand-tuning. We present a study evaluating human efficiency and precision at this task in Section 1.3.2. Even with a simple, easy-to-use GUI optimized to minimize during modal analysis, the study shows that significant human effort is needed to select accurate parameters. The study also finds that humans are able to distinguish between sounds with minor differences in material parameters, suggesting that material parameters from a library may not sufficiently reproduce the sound of a specific real-world object.

Figure 1.1: A real-time interactive virtual environment where striking objects produces dynamic sounds using our method (left); a ball striking plates of various sizes plays a melody (middle); and a set of wind chimes blowing in a virtual forest (right).

[1]This chapter previously appeared as a paper in the 26th IEEE Conference on Virtual Reality (IEEE VR 2019) and an article in a special issue of IEEE Transactions on Visualization and Computer Graphics (TVCG). The original citation is as follows: Sterling, A., Rewkowski, N., Klatzky, R. L., and Lin, M. C. (2019). Audio-material reconstruction for virtualized reality using a probabilistic damping model. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1855–1864

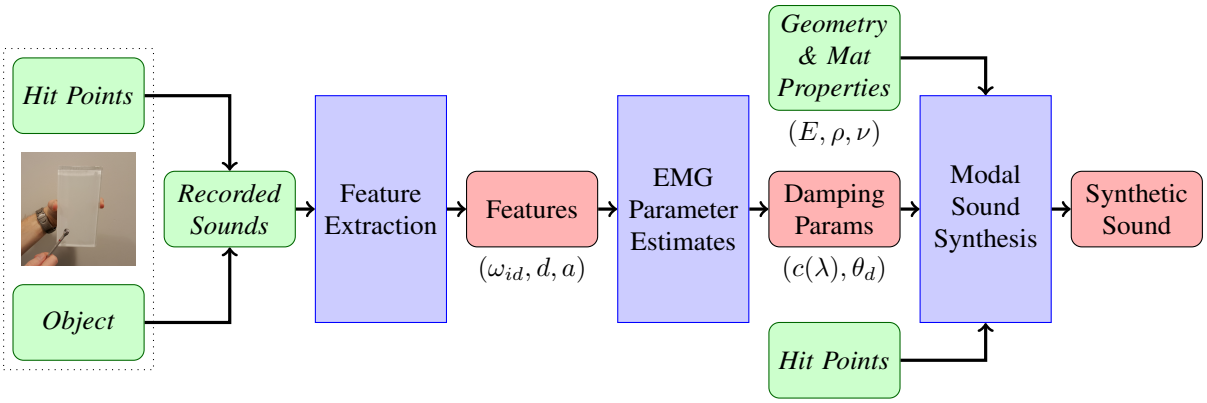

Figure 1.2: Our pipeline for estimating material parameters from recorded audio and using the parameters to synthesize sound for objects of the same material. Inputs are in green with italic text. If the object and hit points are unknown, the pipeline can begin with recorded sounds instead.

Automated material parameter estimation provides a means to estimate the material parameters of a specific object while reducing required human effort. Given an object made of a particular material, we can strike the object and record the resulting sound. Existing methods use the sound, along with mandatory knowledge about the shape and properties of the struck object, to estimate a number of material parameters (Ren et al., 2013b). The material parameters can be applied to sound synthesis, "virtualizing" the audio characteristics of a given material. While recent techniques have been able to estimate material damping parameters, they assume minimal effect on damping from external factors.

For example, an object struck for the purposes of recording either needs to be held by hand or left to rest on another surface. The interface between the object and its *support* will introduce additional damping. To account for this support damping, recordings must be made with supports that introduce minimal damping, requiring a carefully controlled recording environment using special support (Pai et al., 2001), e.g. strings or rubber bands, to suspend the object (Ren et al., 2013a). Other factors that affect estimated damping values, such as complex modes of vibration, background noise, and accumulated error during estimation are assumed by prior work to be minimized. Satisfying all of the assumptions made by prior work requires significant human effort.

In this chapter, we present a practical and efficient probablistic algorithm to estimate material damping parameters directly from recorded impact sounds that accounts for these different factors affecting damping, reducing their effects on the estimated parameters. Unlike previous work (Ren et al., 2013a), this method is fast and requires no prior knowledge about the recorded object's geometry, size, or hit location(s). We are able to virtualize the specific material of a given object. Our method requires significantly less human time

and effort to acquire material damping parameters than previous methods, while producing parameters of similar quality. The key contributions of this work include:

- A new probabilistic material damping model that independently considers each source of damping (Section 1.2.5);
- Application of this probabilistic model to estimation of material damping parameters (Section 1.2.6);
- A study evaluating human effectiveness at manual estimation of material parameters from sound (Section 1.3.2); and
- Quantitative (Section 1.3.3) and perceptual (Section 1.3.4) evaluation of estimated damping parameters.

We validate our method through comparison between estimated and ground-truth damping values, an auditory perceptual study, and comparison against alternative techniques. Figure 1.1 demonstrates our system in several complex virtual environments consisting of real-time interaction with virtual objects of different materials. Figure 1.2 shows the full pipeline for estimating material parameters and using them to synthesize sound.

1.2 Probabilistic Damping Modeling

In order to perform the modal sound synthesis process described previously in **??**, we need to know the object's geometry, Young's Modulus E, density ρ, Poisson's Ratio, and damping parameters α_j for a chosen damping model. We now consider how this information can be obtained in the first place. The geometry can either be taken from a real-world object or designed for a virtual object. Young's Modulus, density, and Poisson's Ratio can be measured from real-world objects, but for many materials these values have been published and approximate values can be selected for synthesis purposes. Damping parameters, on the other hand, are specific to their damping model and are difficult to find for arbitrary materials. In this section, we present our probabilistic damping model for observed damping rates in the presence of external environmental factors.

1.2.1 Hybrid Damping Model

Our probabilistic model extends any of the traditional damping models described in **??**. For this work, we consider the Rayleigh and Caughey damping models previously introduced. We also consider one additional

damping model, derived from generalized proportional damping (see **??**). This hybrid model incorporates Rayleigh damping and a power law damping model (**??**). The damping rates are described according to the function:

$$c_i = \alpha_1 + \alpha_2 \omega_{in}^{2\alpha_3}. \tag{1.1}$$

When α_1 is 0, this becomes the power law damping model, and when α_3 is 1, this becomes the Rayleigh damping model. Since we have found that the optimal damping model varies depending on the object (**??**), this hybrid model can model damping best represented by Rayleigh or power law damping. Using these deterministic damping models, we can now introduce our probabilistic model.

1.2.2 Feature Extraction from Audio

Our technique uses multiple recorded impact sounds to estimate material parameters. A mode that is heavily damped by external factors in one sound may be relatively undamped in another, providing additional information about the range of possible damping values. As damping parameters are geometry-invariant (Ren et al., 2013a) for simple objects often present in virtual enviroments, we do not need to know the object's geometry, its size, or its hit location.

The first step in our approach is to extract the modal components of each input sound. Assuming the sounds come from rigid objects, the sound produced will be mostly modal and can be decomposed into a set of *features*. Each feature corresponds to one mode of vibration and can be parameterized as a damped sinusoid with a damped frequency ω_{id}, an initial amplitude a_i, and an exponential damping coefficient d_i.

We adopt a feature extraction process that identifies likely features, then performs local optimization. This is derived from the feature extraction step of Ren et al. (Ren et al., 2013b), which is described in **??**. For this work, we again adopt the same feature extraction step, but decouple it from the subsequent Match Ratio Product parameter estimation (**??**).

We further extend this standalone feature extraction step to improve robustness. As an additional step, we remove features with d_i under a threshold. These low-damping features are likely to be a constant pitched background noise unrelated to the impact sound. We also remove features below an amplitude threshold, as they are more susceptible to noise. The extracted (ω_{id}, a_i, d_i) features can be converted into pairs of (λ_i, d_i) values by inverting the process in **??**:

$$\lambda_i = \omega_{in}^2 = \omega_{id}^2 + d_i^2. \tag{1.2}$$

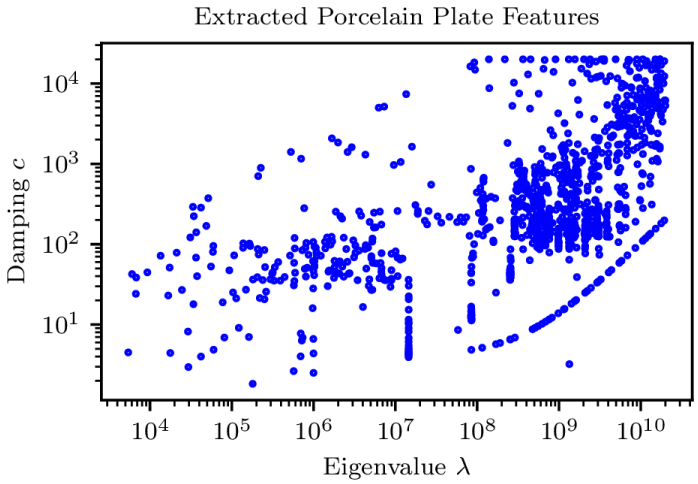

Figure 1.3: Features extracted from multiple impact sounds on a porcelain plate. λ is the eigenvalue of the mode of vibration (related to the frequency), while c is the rate of exponential decay. For any value of λ, there is a range of possible d values, which can be captured in a statistical model.

As a result of this feature extraction process, we have a set of features roughly corresponding to the modes of vibration of the object. The most notable modification is that we account for background noise (modeled as additive white Gaussian noise) by estimating the amplitude of the noise floor. The extracted (ω_{id}, a_i, d_i) features are converted into pairs of (λ_i, c_i) values, where λ_i is the eigenvalue corresponding to that mode of vibration and $c_i = 2d_i$.

1.2.3 Distributions of Damping Values

With (λ_i, c_i) features extracted from multiple input sounds, we now interpret the results. Figure 1.3 shows an example of features extracted from impact sounds on a porcelain plate. Note that for any given eigenvalue λ, there exists a range of extracted damping values. This is especially noticeable where feature points appear as a vertical line, showing that even the same mode of vibration may have a variable rate of decay. These results are inconsistent with the damping models in **????**, which propose a one-to-one mapping between λ and c. Instead, we propose that there is significant error present in the extracted damping value of each feature, and that error can be modeled with a statistical distribution.

The prior work of Ren et al. (Ren et al., 2013b) estimates damping parameters using a least-squares metric to compare spectrograms. Similar results could be produced by fitting a damping model to (λ, c) features using least-squares (e.g. by optimizing all α_j). Statistically, a least-squares fit of a damping model is

equivalent to assuming there is normally-distributed error around the model. We will refer to least-squares fitting of damping models as LSQ. However, we have found experimentally that least-squares fits tend to overestimate the material damping parameters, and resynthesized sounds all sound heavily overdamped.

Another notable property of Figure 1.3 is the clear line of points forming a lower bound to the data (with a few outliers). We have found experimentally that a damping model fit to this lower bound curve results in resynthsized sounds much closer to the input sounds. If the damping model should fit the lower bound, then all error is positive and can only increase the extracted damping values. Statistically, this indicates a one-sided error distribution; e.g. half normal, exponential, or chi-square distribution. Computationally, a lower bound Rayleigh damping model can also be found as a line along the lower convex hull of the points. We refer to lower-bound fitting of damping models as LB.

However, as can be seen in Figure 1.3, outliers often appear below the clean LB curve, and for other objects such a clean curve does not appear in the first place. A strict LB fit will be highly sensitive to outliers, as it must assume all error is positive. It is difficult to detect and remove outliers in extracted feature datasets. To solve this problem, we examine the physical sources of error in extracted damping values and construct an appropriate statistical distribution modeling that error. Ideally, this should produce a more robust lower bound fit which handles outliers based on their statistical probability of occurrence.

1.2.4 External Damping Factors

To accurately model error in damping values, we consider a number of physical phenomena that may affect estimates of the material damping values. These *external* damping factors are distinct from the material damping, which occurs due to the internal structure of a material.

1.2.4.1 Support Damping

An object's method of support can be varied; the object could be sitting on a desk, held in a hand, or dangling from a ceiling. We define a *support* broadly as any long-lasting contact with the sounding object of interest, with enough friction to maintain its contact with the object even when the object is struck. Regardless of the form of support, some energy from the object's vibrations will be transferred to the support, causing additional damping. In real-world situations where the object is unlikely to be minimally supported, the additional damping significantly affects the sound.

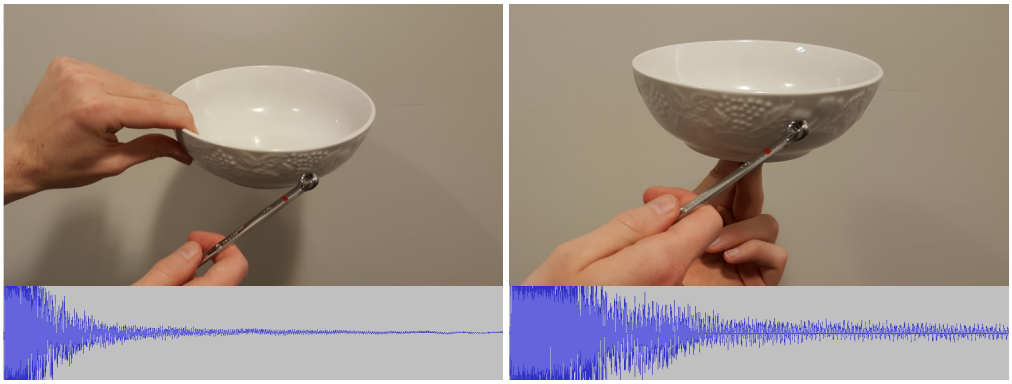

Figure 1.4: A porcelain bowl struck in the same location produces different sound when supported with a tight grip (left) or supported by resting on a single point (right). Without accounting for the effect of the support, prior methods would not be able to estimate accurate material parameters from these sounds.

Refer to Figure 1.4 for an example of the effect of the support on the resulting sound. A tight grip on the bowl's rim produces a more damped sound compared to gentle balancing on fingertips.

1.2.4.2 Complex Modes

Complex modes of vibration are slight deviations from normal mode behavior. Unlike normal modes, complex modes are not linearly separable: energy may be transferred between modes while vibrating. A mode that *loses* energy to others will produce higher damping values, while a mode that *gains* energy from others will produce lower damping values. Most systems have only slightly complex modes (i.e. there is little energy transfer), so normal modes are a close approximation (Imregun and Ewins, 1995), but not an exact one. Since we make the assumption of normal modes, the slight transfers of energy are a source of error in damping value estimates.

1.2.4.3 Background Noise

Background noise in recorded sounds is too variable to realistically model. The feature extraction step of the method is designed to specifically extract *modal* features from the sound. This mostly eliminates persistent "hums" which do not match the modal exponential-decay model. We modify the feature extraction method to account for additive white Gaussian noise (Section 1.2.2). This further removes the influence of persistent background noise, though there may still be some remaining Gaussian (normal) error in the spectrograms and their resulting extracted damping values.

Acoustic reflections and reverberations from room acoustics are confounding factors. Without knowing the properties of the room acoustics, we cannot separate the effect of a damping material from the effect of the acoustics. For our model, we still assume minimal room reverberations, but some small sources of transient noise or early reflections may be appropriately modeled by normally distributed error.

1.2.4.4 Feature Extraction Error

The feature extraction step itself is not perfect; some error is introduced in the process. For example, spectrograms have limited spectral and temporal resolution, and the Fourier transform's assumption of periodicity in each window is an approximation. The discretization of the spectrogram will produce small amounts of error. Sidelobes resulting from Fourier transforms may appear as separate peaks or affect the estimated damping rate of nearby modes.

1.2.4.5 Acoustic Radiation

Uneven acoustic radiation from the object may mean that different microphone placements will result in different initial mode amplitudes. This can be accounted for by keeping the object and microphone stationary during an impact sound. However, the relative positions of the microphone and object do *not* need to be fixed across all input sounds. Moving the microphone between sounds will not change the frequencies or exponential rates of decay, and thus does not need to be accounted for in our model.

1.2.4.6 External Factor Summary

Current damping parameter estimation techniques do not explicitly consider these factors, instead attributing all damping to the material (as we do in **??**). The resulting damping parameters therefore model the combined effect of the material *and the recording environment*. These parameters may not properly transfer to an object of the same material in a *different environment*. This limits the sounds that can be used for accurate damping parameter estimation: the sounds must be recorded in a carefully controlled setting. With a thoroughly robust technique that can separately model environmental factors, we can reduce the factors' impact on the estimated parameters. The external factors cannot be fully removed, but reducing their impact may result in more physically-accurate material parameters.

1.2.5 Generative Model for Combined Damping

We now introduce a generative model for sampling damping values. The model defines the probability distribution for an extracted damping value c_i, given the eigenvalue λ_i and a set of parameters θ. θ contains parameters representing both the *material* and the *environment*. The material damping parameters, such as α_1 and α_2, are referred to as θ_d for generality. The model can be written as $p(c_i|\lambda_i, \theta)$, and asks, "given a known material and environment, what is the probability of measuring any particular damping value?"

The value c_i is a damping value obtained from the feature extraction step. In the *absence* of any external factors, c_i would consist only of material damping. To account for the external factors, we model c_i as a random variable based on the sum of normally and exponentially distributed random variables.

1.2.5.1 Normal Distribution

A normal distribution models the effect of some external factors. The normal distribution accounts for (1) energy transfer due to complex modes, (2) small sources of background noise, and (3) error in feature extraction due to spectrogram discretization. We assume that each of these factors are an additive, normally distributed random variable. The sum of these normally distributed factors (c_i^n) is also normally distributed:

$$p(c_i^n|\lambda_i, \theta_d, \sigma) = \mathcal{N}\left(c(\lambda_i, \theta_d), \sigma^2\right) \tag{1.3}$$

The distribution is centered on the damping function c evaluated at an eigenvalue λ with damping parameters θ_d, with a standard deviation σ resulting from the combination of factors.

1.2.5.2 Exponential Distribution

An exponential distribution models the effect of the object's support.

$$p(c_i^e|\eta) = \mathrm{Exp}\,(\eta) = \eta e^{-\eta c_i^e}. \tag{1.4}$$

c_i^e is the resulting exponential damping resulting from the object's support, while η is the rate parameter of the exponential distribution. This distribution is an approximation, but in attempting to create a robust lower bound method, it serves the role of a one-sided distribution fitting to the lower bound of damping values.

Zheng and James defined a model to approximate additional per-mode damping based on contacts with other objects (Zheng and James, 2011). However, a statistical analysis of this model is highly dependent on the distribution of elements of the matrix of eigenvectors $\mathbf{\Phi}$. We are not aware of any prior work that has attempted to statistically model the distribution of eigenvector matrix $\mathbf{\Phi}$ elements, and our own analysis using Kolmogorov-Smirnov goodness-of-fit tests found no probable common distributions. In the absence of a more well-defined model and with the main requirement of a one-sided distribution satisfied, the exponential distribution was selected empirically based on extracted feature data.

1.2.5.3 Exponentially Modified Gaussian

The combined damping value c_i can then be modeled as the combination of (1) the normally-distributed factors c_i^n due to complex modes, noises, and other sources of errors, and (2) the exponentially-distributed factor c_i^e due to the support damping. Assuming that the factors are independent (for mathematical feasibility), they can be formulated as two separate sources of exponential decay of the mode amplitude z_i:

$$z_i(t) = a_i e^{-c_i^n t} e^{-c_i^e t} \cos(\omega_{id} t) \tag{1.5}$$

$$= a_i e^{-(c_i^n + c_i^e)t} \cos(\omega_{id} t). \tag{1.6}$$

The probability density function of the sum of the normal and exponential distributions ($c_i^n + c_i^e$) is the convolution of their individual probability density functions. The resulting distribution is an *exponentially modified Gaussian* (EMG) distribution. EMG distributions have been used extensively in chromatography (Grushka, 1972), but have also found uses in other domains. The probability density function for the EMG is:

$$p(c_i|\lambda_i, \theta_d, \sigma, \eta) = \frac{\eta}{2} e^{\frac{\eta}{2}(2c(\lambda_i,\theta_d) + \eta\sigma^2 - 2c_i)} \operatorname{erfc}(s_i) \tag{1.7}$$

$$s_i = \frac{c(\lambda_i, \theta_d) + \eta\sigma^2 - c_i}{\sqrt{2}\sigma}, \tag{1.8}$$

where erfc is the complementary error function, defined as:

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-y^2} dy. \tag{1.9}$$

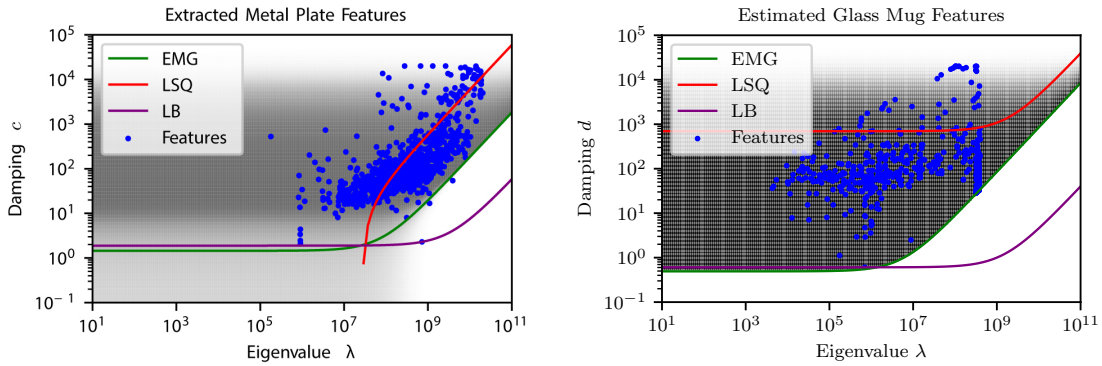

Figure 1.5: Parameter estimation on sound features. Each feature consists of an eigenvalue λ_i and its corresponding damping coefficient c_i. Estimated Rayleigh damping curves are plotted, with the variation from the curve caused by external factors. Our method, using the EMG distribution, provides the closest fit to the lower bound of the data while being relatively unaffected by outliers.

This defines the probability of observing an extracted damping value, given the material damping and environmental damping parameters. This is the complete generative model for damping values, encompassing multiple sources of damping and errors. Since only the frequencies and damping values of the modes are needed for this model, we do not need to assume that the mode shapes remain unchanged. The full set of parameters θ is (θ_d, σ, η), which together define the distribution.

1.2.6 Parameter Estimation

With the generative model established, we now describe the estimation of damping parameters. We estimate the parameters θ through maximum likelihood estimation (MLE). The generative model above uses known parameters to produce data from a distribution. MLE is an optimization method that reverses the process: use known data from a distribution to produce best-fitting parameters. Given a set of extracted (λ_i, c_i) pairs as data and a set of parameters, we can use the generative model to compute the log-likelihood of the data given the parameters:

$$\log p(\mathbf{d}|\boldsymbol{\lambda}, \theta_d, \sigma, \eta) = \sum_i \log\left(\frac{\eta}{2}\right) + \eta c(\lambda_i, \theta_d) + \frac{\eta^2\sigma^2}{2} - \eta c_i + \log\left(\operatorname{erfc}(s_i)\right). \tag{1.10}$$

Using the log-likelihood simplifies computation, removing exponentiation and turning a product of probabilities into a sum of log probabilities. We want to find the parameters that *maximize* this log-likelihood—and hence also maximize the original probability. These maximizing parameters are those that best explain

the extracted data, "fitting" the probability distribution to the data. We compute the analytical gradient of the log-likelihood function and perform gradient ascent to find these optimal parameters.

We compute the full average derivative for the n (λ_i, c_i) samples. We define a term t_i and use the scaled complementary error function $\mathrm{erfcx}(s_i) = \exp(s_i^2)\,\mathrm{erfc}(s_i)$ to simplify notation:

$$t_i = \frac{-2}{\mathrm{erfc}(s_i)\sqrt{\pi}} e^{-s_i^2} = \frac{-2}{\mathrm{erfcx}(s_i)\sqrt{\pi}}. \tag{1.11}$$

The derivatives for η and σ must be computed for all damping models. Their derivatives are as follows:

$$\frac{\partial \log p}{\partial \eta} = \frac{n}{\eta} + n\eta\sigma^2 + \sum_i c(\lambda_i, \theta_d) - c_i + t_i \frac{\sigma}{\sqrt{2}}, \tag{1.12}$$

$$\frac{\partial \log p}{\partial \sigma} = n\lambda^2\sigma + \sum_i t_i \left(\frac{\eta\sigma^2 + c_i - c(\lambda_i, \theta_d)}{\sqrt{2}\sigma^2} \right). \tag{1.13}$$

The derivatives for θ_d will depend on the damping function itself. We will present the derivatives for Rayleigh damping here; derivatives for alternative models are not difficult to compute. For Rayleigh damping's linear $c = \alpha_1 + \alpha_2\lambda$ function, the derivatives for α_1 and α_2 are:

$$\frac{\partial \log p}{\partial \alpha_1} = \eta n + \sum_i \frac{t_i}{\sqrt{2}\sigma}, \tag{1.14}$$

$$\frac{\partial \log p}{\partial \alpha_2} = \sum_i \eta\lambda_i + \frac{t_i\lambda_i}{\sqrt{2}\sigma}. \tag{1.15}$$

With the derivative established, we can perform standard gradient ascent until convergence. The final damping parameters in θ_d are the optimal parameters for the material of the struck object. These damping parameters can be used to represent the recorded material for modal sound synthesis, with other effects (e.g. room acoustics, supports) modeled separately (Zheng and James, 2011).

Section 1.2.6 shows features extracted from 19 impact sounds on a metal plate, while Section 1.2.6 shows features extracted from 40 impact sounds on a glass mug. Similarly, **??** shows features extracted from synthetic sounds. The figure compares our EMG fit with MLE optimization against the baseline LSQ and LB methods (see Section 1.2.3). In each case, LSQ overfits the data, while LB is strongly affected by low outliers and underfits the data.

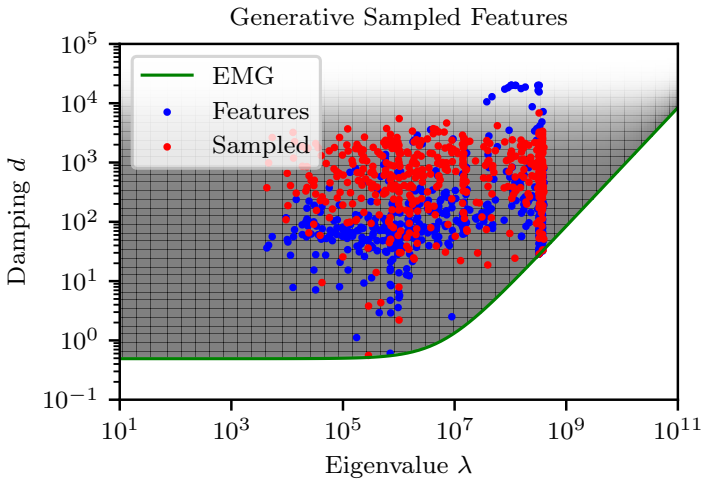

Figure 1.6: Comparison of real-world extracted features (blue) and sampled features from a fitted EMG model (red). The two sets of points are similarly distributed, indicating that the EMG model is properly fit to the real-world data.

1.2.7 Discussion and Analysis

The effect of external damping factors cannot be entirely removed, and in real-world situations the extracted damping values may all be much higher than the material damping function alone. This positively biases the estimator: the estimated parameters will often be larger than the ground truth. By accounting for external factors, this estimator has less bias than other methods, and is therefore more accurate.

Figure 1.6 shows experimental validation of the EMG model. It contains the *same* extracted glass mug features as found in Figure 1.5 in blue, but overlays additional features in red. These red features are sampled from the optimized EMG distribution. They need not correspond one-to-one with the extracted features, but they should follow a similar *distribution*. This shows experimentally that the underlying statistical model is appropriate for capturing the distribution of real-world data.

1.2.8 Sound Synthesis with Estimated Values

The estimated damping parameters should be accurate, having accounted for the effect of the support. However, modal sound synthesis assumes free vibrations (i.e. no support) when in most cases there will be something supporting the object. An additional step is needed to apply support damping to synthesize contact sounds due to support.

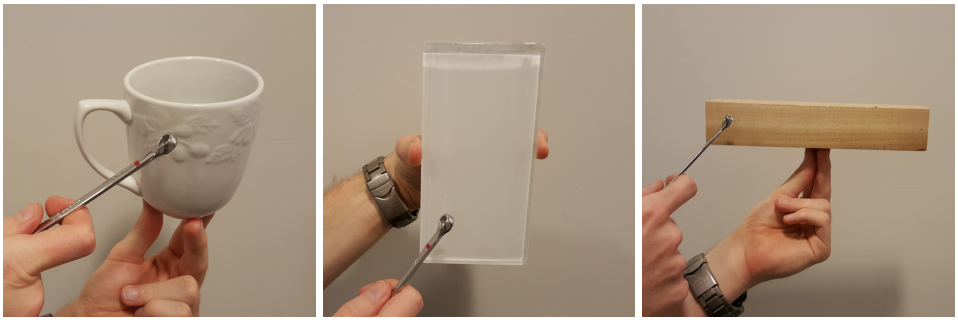

Figure 1.7: Three objects from our impact sound dataset: a porcelain cup (left), a small glass tile (center), and a wood block (right). Note the ways that each object is supported. These supports interfere with damping parameter estimation.

We adopt a contact model for modal sound synthesis introduced by Zheng and James (Zheng and James, 2011). The method uses an additional damping matrix $\mathbf{G}$ to model the additional damping resulting from each contact point k in the set of contact points $\mathcal{C}$:

$$\mathbf{G} = \sum_{k \in \mathcal{C}} c_k \mathbf{\Phi}_k^T (\mu \mathbf{I} + (1 - \mu) \mathbf{n}_k \mathbf{n}_k^T) \mathbf{\Phi}_k, \tag{1.16}$$

where c_k is the magnitude of the contact force for contact k, $\mathbf{\Phi}_k$ is the set of eigenvectors corresponding to the point at contact k, μ is the coefficient of friction, and $\mathbf{n}_k$ is the normal direction at that point. Some mode coupling is introduced since $\mathbf{G}$ is not diagonal, but this coupling was found to be perceptually minor. Therefore, each damping model may be augmented by adding the corresponding diagonal component of $\mathbf{G}$.

1.3 Results

We have implemented the damping parameter estimation method described in Section 1.2.5 and tested its effectiveness through both numerical analysis and perceptual validation. With this method, the process for material damping parameter estimation involves striking an object repeatedly, ideally with varying hit locations and support methods. This approach has less strict requirements about the recording environment than previous work; sounds can be recorded in a quiet room, as long as there are few transient sounds and the room is not heavily reverberative.

We have recorded numerous impact sounds on a set of fifteen rigid objects, where the hit points and the method of support are documented for each impact. Figure 1.7 shows a sample of these objects, with

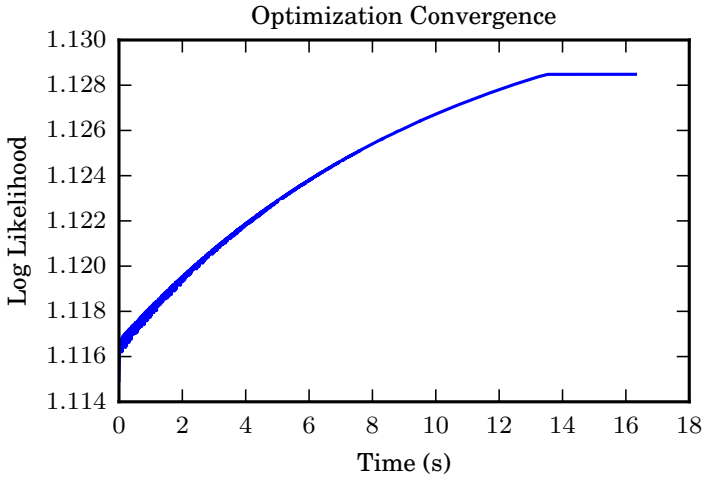

Figure 1.8: Plot of log-likelihood maximization converging over the course of parameter estimation. Optimization was performed on 752 frequency-damping points extracted from porcelain plate impact sounds, and converged after 39,009 iterations for a total of 16.3s in an one-time preprocessing.

various hit locations and methods of support. There are an average of nearly 50 impact sounds sampled per object. All objects were supported by hand, often either with an edge being pinched between two fingers or the center resting on a few fingertips. Audio was recorded using a Zoom H4 in a padded sound recording booth which reduced, but did not eliminate, acoustic effects and background noise. Objects were struck with a small metal wrench, the wrench itself being tightly gripped to minimize its own vibrations.

The eigenvalues and damping values are each normalized, but the data are not shifted or centered. With this normalization, the estimated damping values need to be unnormalized for application to other materials. Although we cannot guarantee that the optimization problem in this context is always convex, especially for higher order damping functions, in multiple runs from different starting points on multiple datasets, all optimization processes have converged on the same parameters. The optimal value of σ tends to be very small, indicating that the distributions of damping values tend to be closer to exponential distributions than to normal distributions. σ and η are used to guide optimization of the damping parameters, but they are not needed for sound synthesis.

We implemented the parameter optimization algorithm in Python and NumPy. On a laptop with a dual core 2.53 GHz Intel Core i5-540M processor, optimization over thousands of features from tens of input sounds and hundreds of thousands of iterations takes 1-5 minutes to complete. See Figure 1.8 to see an example of convergence behavior. Note that we are attempting to *maximize* the log-likelihood, as the

		Porcelain	Travertine	Wood	Steel	Plastic	Glass
Rayleigh	α_1	3.9	1.3	39.0	2.3	39.8	2.0
	α_2	1e-8	2.5e-8	1.3e-7	6.9e-8	1.3e-7	7.8e-8
Hybrid	α_1	3.9	1.3	39.0	2.4	34.83	1.9
	α_2	5.2e-9	2.5e-8	2.1e-7	5.5e-8	4.1e-7	1.5e-7
	α_3	1.027	1.001	0.978	1.011	0.95	0.974

Table 1.1: Damping parameters estimated using our technique. These materials come from a subset of objects in our impact sound dataset. These parameters are described in **??** and Section 1.2.1. When hybrid $\alpha_3 = 1$, the remaining hybrid damping parameters are equivalent to their Rayleigh damping counterparts. These parameters can be used to virtually recreate the material of the real-world object.

parameters that maximize the log-likelihood also maximize the underlying probability. Upon convergence, the optimized θ_d parameters model the damping behavior of the recorded material.

Table 1.1 contains results from estimation on some of the objects. When hybrid $\alpha_3 = 1$, the model is identical to Rayleigh damping. Even small changes in hybrid α_3 can have a large impact on the resulting damping. For example, a 10 kHz mode on the Porcelain plate has a damping coefficient $d = 20$ with the provided parameters (hybrid $\alpha_3 = 1.027$), but changing α_3 to exactly 1 reduces the damping coefficient to $d = 12$.

In general for these damping models, larger parameters create virtual materials with more damping and shorter sounds. For example, the two objects with the most damping are the wood block and plastic bowl, whose materials are known to be naturally heavily damped. The porcelain plate, travertine tile, and glass tile all had similar estimated parameters.

1.3.1 Real-time Synthesis and Rendering

Finally, the optimized parameters are used for sound synthesis. Each sounding object must be preprocessed before running any interactive application. Preprocessing time depends primarily on the number of tetrahedra in the input mesh; a mesh with 2,000 tetrahedra takes under a minute to preprocess while a mesh with 30,000 tetrahedra can take many minutes. Once each sounding object has been preprocessed, modal synthesis is performed in real time at 44 kHz.

Like previous work (Ren et al., 2013b), we are able to synthesize sound using an interactive rigid-body physics simulation in real time. We have implemented our method for sound synthesis with support damping in C++ as a module for Unreal Engine 4. Our demos have been integrated with an HTC Vive headset and

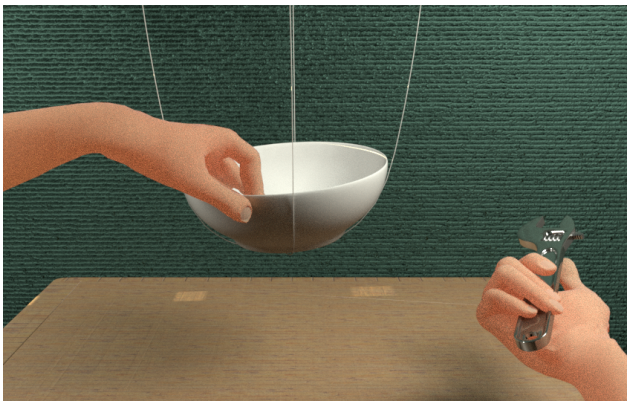

Figure 1.9: A simulated porcelain bowl is struck in multiple locations, with and without a supporting grip.

Leap Motion controller. The user's hands were tracked with the Leap Motion, with the Vive controllers used to represent tools that could be picked up and used to strike objects. Users can walk in the virtual environment and strike objects, immediately hearing the resulting synthesized sound. Figure 1.1 shows multiple scenes from our real-time demo, with multiple objects of various shapes and materials. Figure 1.9 shows another scene, where a bowl is supported by either strings or a hand, producing different sounds depending on the hit point and support type.

1.3.2 Human Hand-Tuning Evaluation

In the absence of an automated method for damping parameter estimation, parameters have traditionally been estimated by hand. We present a study evaluating the effectiveness of human damping parameter estimation, using human subjects to hand-tune material parameters for multiple objects. Specifically, we are interested in the tuning of the damping parameters and the specific modulus γ, defined as the ratio of Young's modulus to density. We seek to evaluate the distributions of subjects' selected material parameters. For example, are subjects able to agree on a single unique set of material parameters, and if so, to what degree of precision? We also seek to determine the time and sound samples needed for subjects to reach their conclusions.

1.3.2.1 Experimental Setup

We constructed an easy-to-use GUI enabling interactive adjustment of material parameters for sounds produced through modal sound synthesis. For each object in the study, we created a corresponding 3D model by hand (a laborious process requiring precise measurements) and performed modal analysis on that model

once (a few hours of computation time). The damping parameters and specific modulus γ for an object can be adjusted as a post-processing step, without needing to repeat the lengthy modal analysis step. With these optimizations, resynthesis with modified parameters took less than two seconds, allowing for rapid iteration. In the interface, each parameter was controlled with a slider, with a range of plausible realistic values presented on normalized scales from 0–100.

Subjects were recruited primarily through mailing lists and were not required to have any background in parameter tuning or impact sound analysis. Subjects were compensated financially for their participation. Subjects were given real-world objects, placed on small foam blocks to reduce support damping. For each object, the subjects' task was to tune material parameters such that the synthesized sound produced by the application most closely matched the sound they heard when striking the real object. Subjects were instructed to find the most accurate parameters possible, regardless of the time needed. Subjects first performed this task with a training object, in order to reduce learning effects. The six objects evaluated were all disk-shaped objects of approximately the same radius and thickness. Every subject hand-tuned material parameters for all six objects in a random order.

The study was divided into two sections. The first 20 subjects hand-tuned three parameters for each object: the two Rayleigh damping parameters and the specific modulus. The following 20 subjects hand-tuned two parameters for each object: just the two Rayleigh damping parameters. For the two-parameter section, the specific modulus was set to the mode of the subject-selected specific moduli from the three-parameter section. The three-parameter section models the real-world case where all three parameters must be picked in order to virtualize an object.

1.3.2.2 Results and Analysis

We first consider the distributions of subjects' selected material parameters. Figure 1.10 shows results from the three-parameter section of the study for a few selected objects: a wood disc and a porcelain disc. For highly reverberant objects such as the porcelain disk, subjects could generally agree on Rayleigh damping's α_1 parameter for each object. However, for highly damped objects such as the wood disk, Rayleigh α_1 responses were less consistent. The distributions of Rayleigh α_2 parameters for each object show agreement between subjects, indicated by the relatively low standard deviations and frequently unimodal distributions. The specific modulus, which modifies the pitch of the synthesized sound, often resulted in multimodal distributions.

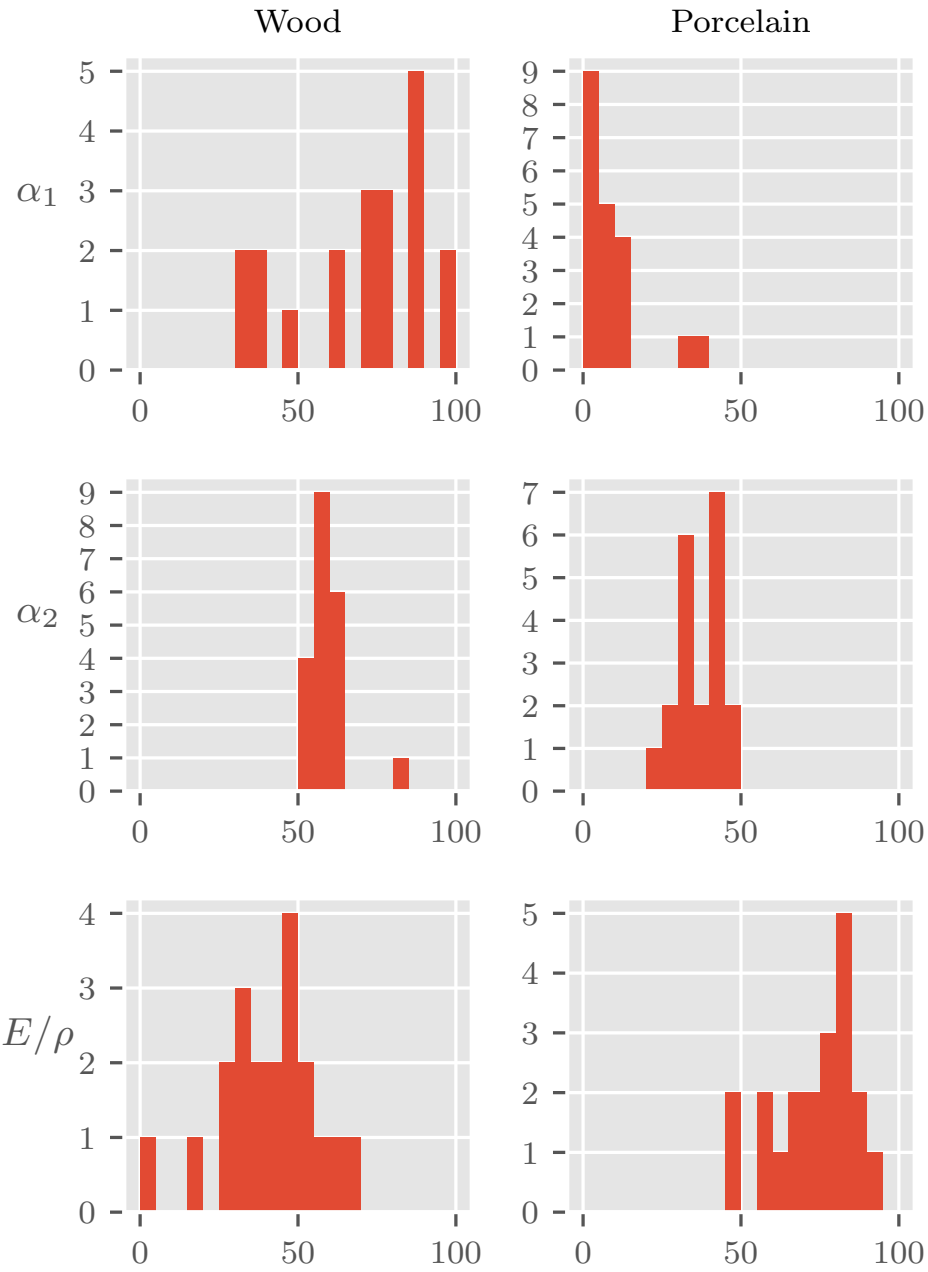

Figure 1.10: Distributions of human-tuned material parameters for wood and porcelain discs. α_1 and α_2 are Rayleigh damping parameters, E/ρ is the specific modulus, and all parameters were tuned on normalized scales from 0–100. The observed distributions indicate that subjects had difficulty finding a unique optimal solution for the specific modulus, and for α_1 in the case of highly-damped objects.

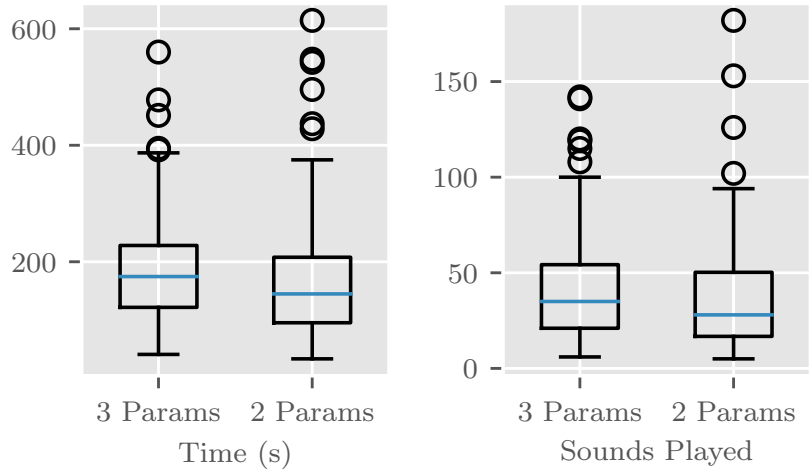

Figure 1.11: Box-and-whisker plots of the time and number of synthesized sounds needed for subjects to reach their final hand-tuned material parameters, with deviant observations plotted as outliers. Between the two versions of the study, the difference in median time is 20 s, and the difference in median sounds played is 7 sounds. Our method is an automated version of the 2-parameter study, and significantly reduces the human labor needed.

We also consider the time and number of sounds needed for subjects to reach their conclusions. Figure 1.11 contains histograms for the amount of time and number of synthesized sounds needed for subjects to finalize their selections. The median time needed was 165 s to tune three parameters, and 145 s to tune two parameters. The median number of synthesized sounds needed was 35 to tune three parameters, and 28 to tune two parameters. The range of times (33–614 s) and sounds (5–182) was highly variable, and the effect of parameter count on times and sounds did not reach statistical significance by t-test.

Our method for parameter estimation is an automated way to perform the parameter selection task. Human hand-tuning requires around 145 s and 28 sounds per object, requiring dedicated human attention for the entire duration. Hand-tuning also requires creating an accurate 3D model of the object and performing modal analysis, possibly adding hours of extra human effort. In contrast, our automated method operates effectively with 10–20 sounds and does not require a 3D model of the object. Parameter estimation then takes a few minutes, during which no human attention is required. Overall, our method significantly reduces the amount of human labor needed to create virtualized objects. Compared to prior work, our method reduces human labor by not requiring carefully controlled recording environments, creation of a 3D model, and knowledge of object geometry and hit points.

1.3.3 Synthetic Validation

Synthetic validation provides a numerical comparison against ground-truth damping parameters. We synthesized a variety of sounds with known damping parameters and passed the resulting sounds through the parameter estimation process to see if the original ground-truth values could be recovered using our algorithm. Sounds were synthesized from the geometry of 18 models, ranging from small, hollow cups to desktop vases and large sculptures. Five materials were chosen by randomly sampling material parameters from a range of realistic values. For each object, ten support points were sampled at random on the surface of the object, each with a random amount of contact force ranging from a light support to a moderate support. Then, 100 sounds were synthesized for each combination of object and material. Each sound sampled its impact point randomly on the exterior surface and picked one support point to be active. The resulting sounds were passed through the feature extraction process for Rayleigh damping, and extracted features from a varying number of sounds were used to estimate the original parameters.

Parameter estimation was performed with three different estimators: EMG (our method, see Section 1.2.5) and the two baselines LSQ and LB (see Section 1.2.3). Direct comparison against the algorithm of Ren et al. (Ren et al., 2013b) is infeasible due to the significant differences in inputs and outputs. However, their method will produce results most similar to the least squares (LSQ) estimator. We compared the error between the ground-truth parameters and the estimated parameters while using a varying number of input sounds. For each tested number of impact sounds, 30 different sets of sounds of that cardinality were sampled, and the resulting errors averaged.

1.3.3.1 Discussion

Figure 1.12 shows the relative error for each parameter and each estimator. For all materials in this synthetic data, both the EMG and LB estimators significantly outperformed the LSQ estimator ($p < .05$). With real-world data, the EMG and LB estimators more frequently decouple, as the EMG estimator's statistical model better adapts to noise and other artifacts of recording. These synthetic sounds, without noise or other effects, are the ideal situation for the LB estimator, and do not leverage the full capabilities of the EMG estimator.

Rayleigh α_1 estimates have minimal error, especially with larger amounts of data. While the error in Rayleigh α_2 is relatively higher, no prior work has performed a similar validation for comparison. Prior work

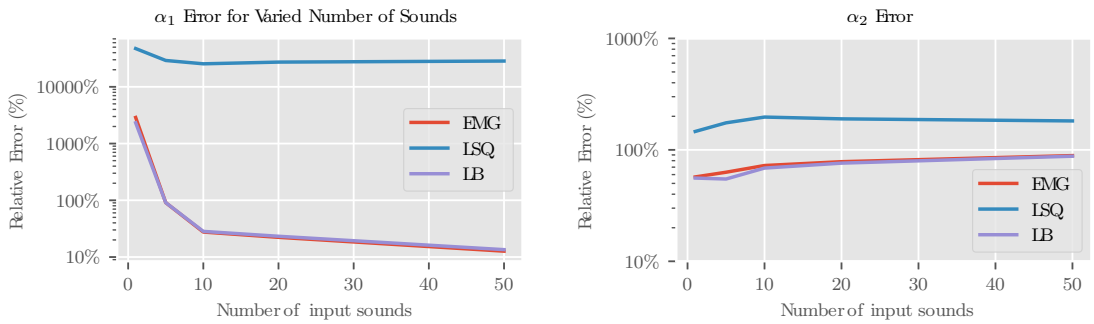

Figure 1.12: Relative error for Rayleigh damping parameters α_1 and α_2 in synthetic validation. As the number of sounds used for parameter estimation increase, error in Rayleigh damping parameter α_1 decreases while α_2 displays some overfitting as number of input sounds increases.

would produce results most similar to the LSQ estimator, which was outperformed by our method. In light of this, our method provides an improvement over previous work while removing the need for knowledge of geometry and hit points. Finally, the error is mostly important as it affects users' perception of the material. Our perceptual evaluation provides an analysis of whether our estimated parameters are accurate when evaluated by humans.

1.3.4 Perceptual Evaluation

Numerical comparisons against previous work are difficult since our method is the first work to estimate damping parameters given only input audio with no knowledge of geometry, size, or hit point. In this study, we considered recorded real-world sounds, and sounds synthesized using three sets of damping parameters: parameters from Ren et al. (Ren et al., 2013b), parameters from the human hand-tuning study (Section 1.3.2), and parameters estimated using our method to create 4 datasets. We sought to evaluate how well the synthesized sounds recreate the real-world sounds. Subjects evaluated sounds individually, answering questions about qualities of the sounds and estimating properties of the object or impact. Synthesized sounds that more accurately recreate qualities of the real-world sounds should produces similar patterns of answers to questions.

1.3.4.1 Experimental Setup

The study was conducted in an online web questionnaire, and subjects were recruited through mailing lists and online posts, but no financial compensation was offered. No prior experience in auditory perception

was expected. Subjects were asked to wear headphones or earbuds to ensure a consistent auditory environment. All sounds were scaled to the same volume, though difference in sound playback devices may have affected perception. Subjects listened to a series of impact sounds, answering questions about each. Variables involved are sound datasets (4: as listed above), object shape (2: disc or rod), and material class (5: wood, metal, plastic, glass, porcelain). All together, this produces a total of 40 sounds to evaluate.

24 subjects participated in the study, but more specific demographic information was not collected. Each subject listened to all 40 impact sounds in randomized order. For each sound, subjects were asked which object shape and material class they suspected created the sound. Subjects also were asked to rate descriptive qualities of the sound—the duration, ringiness, tonality, and pitch—on 7-point ordinal scales. The extreme ends of the scales were descriptively labeled, e.g. tonality ranged from "mixed tones" to "pure tone". Subjects could listen to each sound multiple times as needed. A brief training section at the beginning provided example sounds and definitions for the descriptive qualities.

1.3.4.2 Results: Confusion Matrices

Even with recorded real-world sounds, user identification of material and shape is not always accurate. In evaluation of synthetic datasets, we compare the pattern of errors to those of the real-world sounds, with a closer match suggesting more realistic synthesized sounds. Figure 1.13 shows confusion matrices for material class identification for the disc-shaped objects.

The recorded dataset demonstrates mis-labelings such as heavy confusion between wood and plastic, perception of the glass and ceramic discs as metal, and high accuracy on the metal disc. The hand-tuned dataset differs primarily in perception of its glass and ceramic objects; these differences could be due to human error while hand-tuning or due to inherent assumptions in the underlying modal synthesis model. The Ren dataset largely reproduces the matrix from their original paper (Ren et al., 2013b), although it does not recreate the error patterns (particularly metal) seen in our recorded or hand-tuned datasets. Our dataset (EMG) most closely resembles the hand-tuned results, with the exception of plastic being identified as ceramic by some subjects.

We evaluate the pairwise similarity between these matrices by computing the Frobenius norm of the element-wise difference of the two. The two most similar matrices are the hand-tuned and EMG dataset results, with a difference norm of 16.03. In comparison, between Ren and the hand-tuned data, the norm is 22.09. Against recorded sounds, EMG's norm was 23.11, while ren's norm was 31.85 and hand-tune's norm

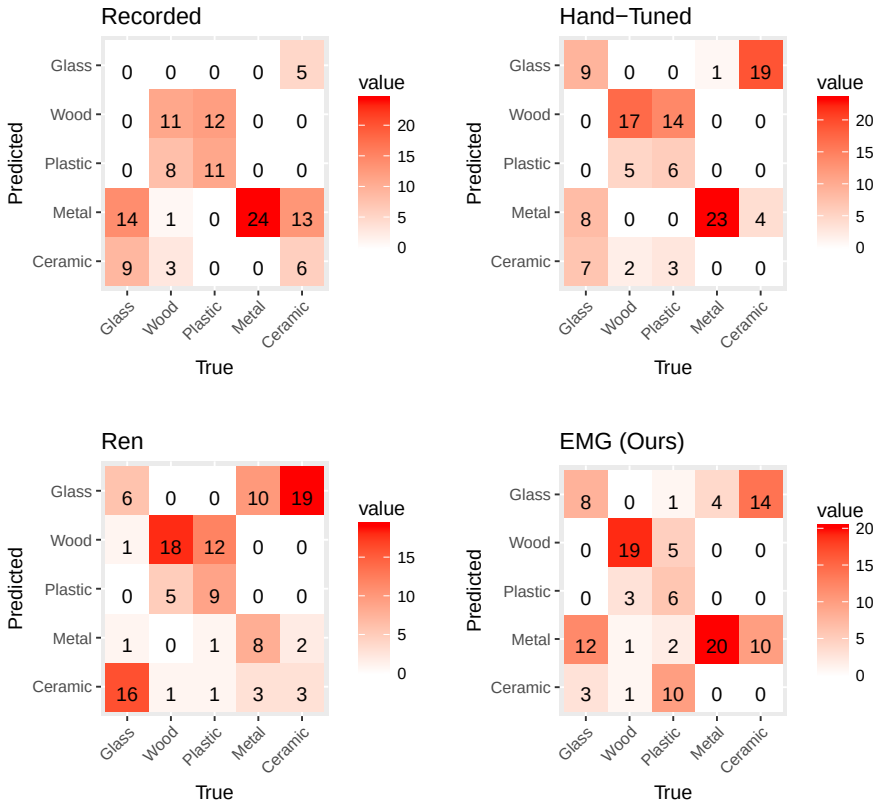

Figure 1.13: Material confusion matrices for the disc-shaped objects in our perceptual study. All four tested datasets (including recorded sounds) show significant labeling errors. However, our method (EMG) replicates the pattern of errors seen in the recorded and hand-tuned datasets more closely than the Ren dataset, suggesting more accurate recreations of the real-world sounds.

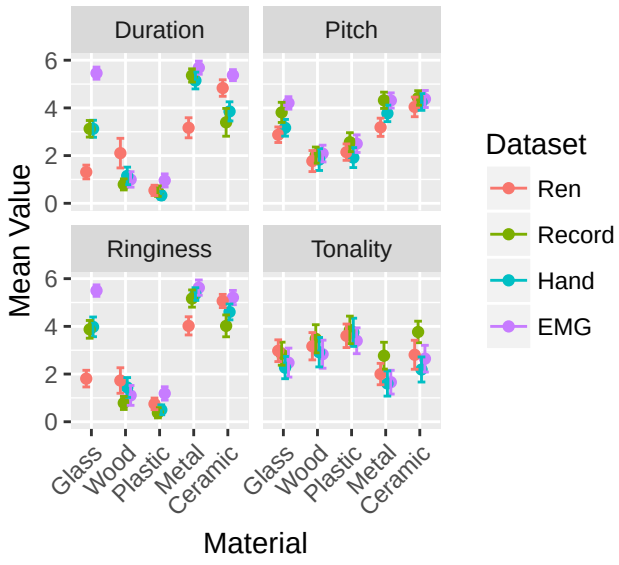

Figure 1.14: The mean selected value for each descriptive quality, material, and dataset. Most datasets are tightly clustered for pitch and tonality, with more differences in duration and ringiness. The Ren dataset contains many statistically-significant differences from the recorded dataset. While our EMG dataset contains some differences in duration and ringiness, it is overall closer to the recorded means.

is 22.83. The high similarity (low difference norm) between our EMG results and the hand-tuned results suggests that our method automatically produces sounds perceptually similar to what would be selected by human hand-tuning.

1.3.4.3 Results: Descriptive Qualities

We evaluate the descriptive quality ratings by performing a multi-factorial repeated measures ANOVA. Each of the variables (sound dataset, object shape, and material) is considered an ANOVA factor, each a repeated measure across subjects. The main effects of dataset, shape, and material are all significant for all four descriptive qualities. For example, for perception of pitch, the effects of material ($F_{4,92} = 95.61, p < .05$), shape ($F_{1,23} = 30.35, p < .05$), and dataset ($F_{3,69} = 22.79, p < .05$) are all significant. This is not surprising, as each of these effects alone can dramatically change the sound. Almost all interaction effects are significant, with the exceptions of material*dataset on tonality ($F_{12,276} = 2.814, p = .107$). Table 1.2 contains the full list of main and interaction effects for each perceptual quality.

Figure 1.14 contains the mean values for each descriptive quality, material, and dataset (using combined results for shapes). For pitch and tonality, most dataset means are closely-clustered. Duration and ringiness

	Df1	Df2	F value	$Pr(> F)$
Duration				
Material	4	92	739.4	$< .001$
Shape	1	23	65.65	$< .001$
Dataset	3	69	81.23	$< .001$
Material*Shape	4	92	52.97	$< .001$
Material*Dataset	12	276	53.56	$< .001$
Shape*Dataset	3	69	43.16	$< .001$
Mat*Shape*Dset	12	276	16.3	$< .001$
Pitch				
Material	4	92	95.61	$< .001$
Shape	1	23	30.35	$< .001$
Dataset	3	69	22.79	$< .001$
Material*Shape	4	92	5.754	$< .001$
Material*Dataset	12	276	2.814	.0012
Shape*Dataset	3	69	69.23	$< .001$
Mat*Shape*Dset	12	276	4.691	$< .001$
Ringiness				
Material	4	92	469.7	$< .001$
Shape	1	23	35.07	$< .001$
Dataset	3	69	51.47	$< .001$
Material*Shape	4	92	40.25	$< .001$
Material*Dataset	12	276	28.45	$< .001$
Shape*Dataset	3	69	7.708	$< .001$
Mat*Shape*Dset	12	276	4.984	$< .001$
Tonality				
Material	4	92	9.325	$< .001$
Shape	1	23	20.97	$< .001$
Dataset	3	69	15.27	$< .001$
Material*Shape	4	92	11.54	$< .001$
Material*Dataset	12	276	1.548	0.107
Shape*Dataset	3	69	35.36	$< .001$
Mat*Shape*Dset	12	276	2.53	.0035

Table 1.2: Significance of fixed effects in our perceptual study, as determined by repeated measures ANOVA. For each of the four perceptual qualities, the resulting degrees of freedom, F score, and p value are listed. All main effects and almost all interactions are significant at the $p < .001$ level.

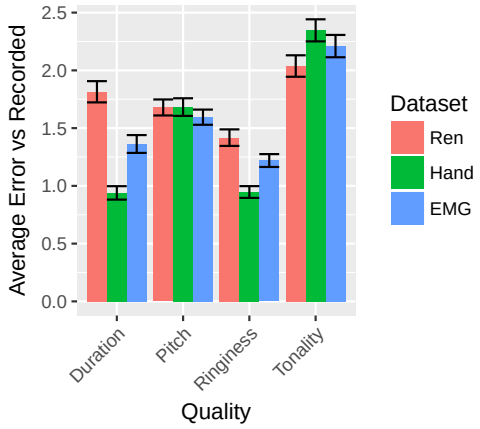

Figure 1.15: Average error between perceptual quality ratings in the recorded dataset versus each of the three synthetic datasets. A lower average error is better, indicating more perceptual similarity with the recorded sounds. Both hand-tuned and EMG perform well.

show more difference: while hand-tuned and recorded are closely aligned, Ren and EMG occasionally display more variance. Ringiness displays nearly the same pattern of significance as duration.

Synthetic datasets that produce more realistic sounds should have descriptive qualities similar to the recorded dataset. To evaluate this, we look at the absolute error between each subject's recorded and synthetic ratings for each sound. Across all materials and shapes, hand-tuned sounds were closest to the recorded sounds for duration and ringiness. For pitch and tonality, all datasets produced more similar results.

Figure 1.15 shows the computed average error values. The hand-tuned dataset is closest for the duration and ringiness qualities, our EMG method is the closest for pitch, Ren is closest for tonality. Our EMG dataset outperforms the Ren dataset on each perceptual quality except for tonality. One-way repeated-measures ANOVAs show a significant effect of dataset on error for duration ($F_{2,46} = 50.03, p < .05$), ringiness ($F_{2,46} = 20.11, p < .05$), and tonality ($F_{2,46} = 5.717, p < .05$). There was no significant effect of dataset on error for pitch ($F_{2,46} = 50.03, p = .347$).

Hand-tuned parameters produce the closest ratings to the recorded dataset on these perceptual scales. The Ren dataset contains many discrepancies from the real dataset, shown in the duration, pitch, and ringiness of the more reverberative materials. Our EMG dataset properly reproduces the perceptions of the pitch and tonality of the recorded objects, but in some cases produces higher duration and ringiness. Our EMG method demonstrates an improvement over the Ren dataset in the fit to recorded sounds.

1.4 Summary

We have presented a method for estimating material damping parameters using recorded impact sounds as the only input. We have validated these contributions through parameter estimation on a new dataset of impact sounds on rigid objects, using both an auditory user study and synthetic validation. These methods can extract real-world material parameters from audio recording and recreate virtualized materials and their rich sound effects arising from dynamic interaction in virtual environments.

1.4.1 Limitations

Our method removes a number of assumptions used by prior damping parameter estimation techniques (Ren et al., 2013b). For example, our method does not require knowledge of the object's geometry, and it reduces the strict assumptions on the object's support and the presence of background noise. However, some common assumptions of prior works remain: (1) application to rigid objects and their vibrations can be accurately modeled by linear analysis. (2) difficulty to fully remove all external damping factors—the presence of loud transient noises, a tightly-coupled support, or a highly reverberative room may still impose residual effects.

1.4.2 Future Work

In general, for parameter estimation, there exists a tradeoff between the amount of assumptions on the required inputs and the quality of outputs. We do not assume prior knowledge on the object geometry, size, material parameters, or the impact location—just audio recording is sufficient. However, this technique currently does not estimate Young's modulus, Poisson's ratio, density, or geometric properties of the object. Generalization of a probabilistic model like this work or use of learning algorithms can potentially estimate these parameters automatically using only a few audio recordings. A method that can optimize all parameters simultaneously would further simplify the pipeline from audio recording to automatic synthesis. Future work may explore if additional inputs can result in a much greater increase in the number of estimated parameters. A single sound is not enough to estimate parameter α_1 with sufficient accuracy; upwards of 10–20 sounds may be needed. In our human parameter tuning study, subjects were initially untrained; experts may produce slightly different parameter distributions.

BIBLIOGRAPHY

Grushka, E. (1972). Characterization of exponentially modified gaussian peaks in chromatography. *Analytical Chemistry*, 44(11):1733–1738. PMID: 22324584.

Imregun, M. and Ewins, D. J. (1995). Complex Modes - Origins and Limits. In *Proceedings of the 13th International Modal Analysis Conference*, volume 2460, page 496.

Pai, D. K., Doel, K. v. d., James, D. L., Lang, J., Lloyd, J. E., Richmond, J. L., and Yau, S. H. (2001). Scanning physical interaction behavior of 3D objects. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 87–96. ACM.

Ren, Z., Yeh, H., Klatzky, R., and Lin, M. C. (2013a). Auditory perception of geometry-invariant material properties. *Visualization and Computer Graphics, IEEE Transactions on*, 19(4):557–566.

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

Sterling, A., Rewkowski, N., Klatzky, R. L., and Lin, M. C. (2019). Audio-material reconstruction for virtualized reality using a probabilistic damping model. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1855–1864.

Zheng, C. and James, D. L. (2011). Toward high-quality modal contact sound. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2011)*, 30(4).