# A PERCEPTUALLY INSPIRED GENERATIVE MODEL OF RIGID-BODY CONTACT SOUNDS

*James Traer*
Dept. of Brain and Cognitive Sciences
MIT
Cambridge, MA, USA
jtraer@mit.edu

*Maddie Cusimano*
Dept. of Brain and Cognitive Sciences
MIT
Cambridge, MA, USA
mcusi@mit.edu

*Josh H. McDermott* *
Dept. of Brain and Cognitive Sciences
MIT
Cambridge, MA, USA
jhm@mit.edu

## ABSTRACT

Contact between rigid-body objects produces a diversity of impact and friction sounds. These sounds can be synthesized with detailed simulations of the motion, vibration and sound radiation of the objects, but such synthesis is computationally expensive and prohibitively slow for many applications. Moreover, detailed physical simulations may not be necessary for perceptually compelling synthesis; humans infer ecologically relevant causes of sound, such as material categories, but not with arbitrary precision. We present a generative model of impact sounds which summarizes the effect of physical variables on acoustic features via statistical distributions fit to empirical measurements of object acoustics. Perceptual experiments show that sampling from these distributions allows efficient synthesis of realistic impact and scraping sounds that convey material, mass, and motion.

## 1. INTRODUCTION

The sounds that enter the ear are collectively determined by the physical processes that generate the acoustic waveform. Sound generation by rigid bodies is a classic physics problem and the processes by which material parameters (e.g. material, mass, motion) affect acoustic waveforms have been well characterized [11, 15, 27, 31]. Typically, physical sound synthesis is done by modelling in detail the relevant processes which lead to the generation of a sound. For example, rigid bodies are modelled as a mesh-grid of masses on springs [4, 5, 6, 28, 38, 41], or decomposed into small segments over which wave equations can be solved by Finite-Element or Boundary-Element-Methods (FEM/BEM) [3, 16, 23]. These models yield a set of resonant modes from which contact sounds can be synthesized. In practice such models require computing physical interactions at very small spatiotemporal scales, and are thus computationally expensive.

Humans perceive sounds in terms of physical variables [12, 34], and these perceptual abilities might inform sound synthesis approaches. When we hear the sound of a fork dropped upon a wooden table, we can make judgments about the size [7, 14, 37], material [2, 13, 17] and motion of the fork [19]. However, our discrimination abilities are limited. It is not clear that humans can tell a fork from a knife in such a case, for instance, let alone the detailed geometry of the fork. Indeed, perceptual experiments indicate that humans can infer broad material differences (e.g. metal vs wood) from contact sounds, but are less accurate for more precise judgments (e.g. distinguishing metal from glass) [13].

The coarse-grained nature of human material judgments suggest material perception is insensitive to mode properties within some tolerance. Exactly what tolerance remains an open question, but it suggests that synthetic modes need not have a detailed correspondence to those of an actual object to yield compelling sounds. We hypothesize that the auditory system infers coarse-grained material parameters from statistical properties of modes, rather than their precise details. For example, consider again the sound of a fork dropped upon a table. Although fine-grained features (e.g. the thickness of the handle, the length of the tines, the narrowing of the neck, etc.) may affect individual modes, we see little evidence that humans infer such subtle features. However, coarse-grained physical features, which are crucial to inferring scene properties like material and size, will affect all the modes and thus are likely to be reflected in the modal statistics.

Rather than attempt to simulate the physical process in fine-grained detail, we measure statistics of modes from real-world impact sounds and use these distributions as the building blocks for sound synthesis via a source-filter model (in which a time-varying force is convolved with the object impulse response). We synthesize sounds from both impacts and sustained frictional forces (Fig. 1). As with our statistical model of modes, the impact forces are parametrized only by coarse-grained properties: mass, stiffness, and velocity. For scraping sounds, the force is generated through a texture quilting algorithm [10], reflecting listeners' perception of summary statistics as opposed to fine-grained temporal detail in sound textures [25].

Our approach yields compelling renditions of sounds via a fast and efficient process. As with other similar approaches [1, 29], it is thus ideal for use in physics engines used in modern computer games and simulations. Such engines store a set of attributes for rigid-bodies to compute how they will move (e.g. mass, elasticity, frictional coefficients, a grid-model of the geometry, etc.) and to compute their appearance under lighting (e.g. diffuse and specular reflectance profiles, visual surface statistics, etc.). As conventional sound synthesis is slow, current engines rely on memory intensive sample banks of pre-recorded or pre-computed sounds to be played on contact. However, our synthesis model only requires a simple texture model and low-dimensional representations of coarse physical features, such as are already encoded for motion and visual appearance. From these crude features and a sample bank of mode distributions (e.g. wood, metal, plastic, ceramic, etc.), our synthesis algorithm can rapidly generate a range of realistic and unique contact sounds. Here we show that impact sounds generated in this

* Work supported by the Center for Brains, Minds and Machines and The MIT-IBM Watson AI Lab

way convey mass and material to listeners as well as recordings of real sounds. Scraping sounds derived from these mode distributions are also realistic and convey motion trajectories.

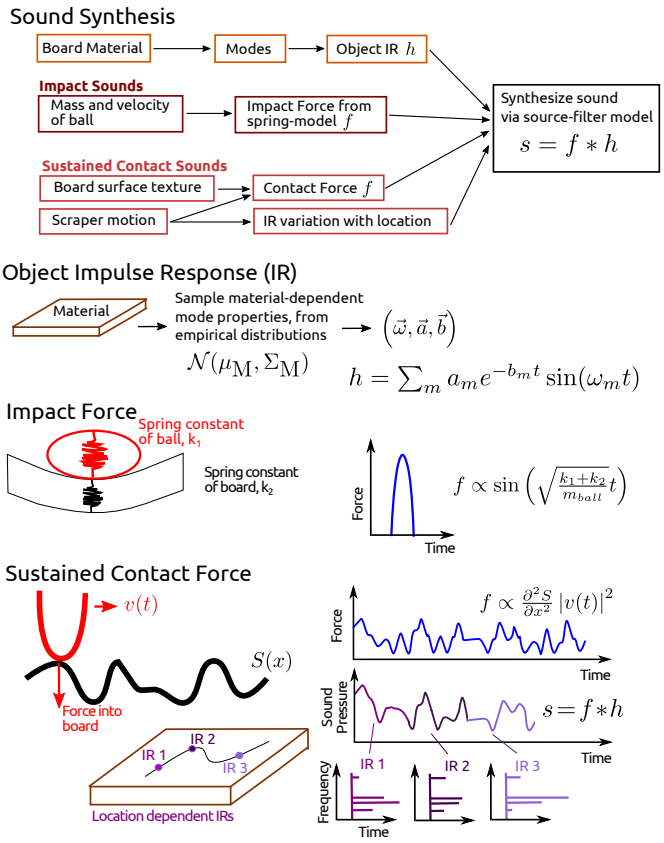


Figure 1: *We synthesize sounds by (top) a generative model of impact and sustained contacts. (Upper-middle) Object Impulse Responses are synthesized by sampling modes from empirical distributions. (Lower-middle) Impact forces are modelled via a spring model. (Bottom) Sustained contacts are modelled via measured surface textures and location-dependent IRs.*

## 2. SOURCE-FILTER MODEL OF IMPACTS

Our model is inspired by the well-known source-filter model [8]

$$s(t) = f(t) * [h_1(t) + h_2(t)] \quad , \tag{1}$$

where $s(t)$ is the sound entering a listener's ear, $f(t)$ is the contact-force between two objects and $h_j(t)$ is the impulse response (IR) of the $j$the object. Past sound synthesis techniques have computed high-resolution IRs with large grid models such as finite-element or boundary-element techniques [16, 24, 30], solved analytically for the resonant modes of an object of known material and shape [17, 21, 22, 38], or fit parameteric models of mode parameters to measured impacts[1]. The grid solutions are flexible but require significant computational power. The analytical modal solutions allow fast synthesis but only apply to a small subset of rigid bodies.

We approximate, as have others before [33], object IRs via the summation of a broadband transient "click" and a set of exponentially decaying sinusoids corresponding to the resonant modes of the object

$$h(t) = h_T(t) + \sum_m^M 10^{(a_m - b_m t)/20} \cos(\omega_m t) \quad , \tag{2}$$

where $h_T$ is the transient, and $(a_m, b_m, \omega_m)$ are the onset power, decay rate and angular frequency of the $m$th mode. The transient can be described via a set of decaying noise-bands:

$$h_T(t) = \sum_n^N 10^{(\alpha_n - \beta_n t)/20} \nu_n(t) \quad , \tag{3}$$

where $\nu_n$ is a time-series of random noise filtered by the $n$th Equivalent-Rectangular-Bandwidth (ERB) filter of a cochleagram decomposition, and $(\alpha_n, \beta_n)$ are the onset power and decay rates of this channel. Under our model, an object IR can be completely described by $2N + 3M$ parameters, to precisely determine the shape of the transient and the modes. Throughout this work we use $N$=30 and $M$=15, which we found to be sufficient for compelling resynthesis.

Our preliminary experiments suggest several broad perceptual trends: (1) perception of material properties is dominated by a small number of powerful modes; (2) changes to the properties of weaker modes are barely noticeable; (3) slight changes to the most powerful modes are detectable, but the resulting sound is perceived as a different exemplar of a similar object or the the same object struck in a different location; (4) altering the transient but not the modes, has a minimal effect on perceived material. All of these perceptual trends suggest that human perception of object properties (i.e. material, size, shape) are primarily predicated upon the statistics of the most powerful object resonant modes.

### 2.1. Modal synthesis of object Impulse Responses (IRs)

To test our hypothesis that human judgments of object properties are based on mode statistics, we seek to synthesize impact sounds which match the modal statistics of real-world impacts, but are otherwise unconstrained (such that the exact mode parameters are different). We began by measuring the mode statistics from real-world objects.

To measure resonant modes, we recorded the sounds of a large number of materials being struck by small pellets. We estimated the resonant modes of each impact via an iterative procedure of spectrogram matching: (1) we obtained the frequency channel of the spectrogram of the impact sound with the maximum power; (2) we synthesized an initial synthetic impact with an exponentially decaying sinusoid at that frequency; (3) we adjusted the mode properties (frequency, onset power and decay rate) to minimize the mean-squared error between the spectrograms of the recording and the synthetic; (4) we subtracted the synthetic spectrogram from the original (removing the mode we just measured). We then repeated the procedure 14 times, yielding parameters for the 15 most powerful modes. After fitting the modes we repeat this procedure using exponentially decaying noise-bands instead of sinusoidal modes to fit the properties of the transient.

For each material, we recorded multiple impacts at different locations on multiple objects. We pooled together modes from multiple objects and characterized the mode statistics by fitting a multivariate Gaussian distribution to the resulting collection. We similarly fit distributions to the transient decay parameters.

To generate a synthetic IR, we sample both mode and and transient properties from our empirically measured distributions:

$$\begin{aligned}(\vec{a}, \vec{b}, \vec{\omega}) &= \mathcal{N}(\mu_{\mathrm{M}}, \Sigma_{\mathrm{M}}) \\ (\vec{\alpha}, \vec{\beta}) &= \mathcal{N}(\mu_{\mathrm{T}}, \Sigma_{\mathrm{T}}) \quad , \end{aligned} \tag{4}$$

where $(\mu_M, \Sigma_M)$ are the mean and covariance of the mode properties, conditioned upon the required object or material, and $(\mu_T, \Sigma_T)$ are the analogous mean and covariance of the transient subband properties. We used rejection sampling to ensure that the average frequency spacing between sampled modes was within 10% of that measured from recordings of the material. Because the mode statistics are computed offline prior to synthesis, all that needs to be encoded at time of sound synthesis are material labels which index distributions of IR properties.

To simulate multiple contacts of the same object we sample from the distributions once, and then randomly perturb mode onset powers (standard deviation=20% mean mode power) for each later impact. This emulates the fact that impacts in different locations differentially excite the same modes. We found empirically that either sampling from the distribution twice or repeating the exact same set of mode parameters produced unrealistic sounds [20].

### 2.2. Effect of impact physics

To synthesize an impact sound, we also need to compute the contact force, to be convolved with the object IR [Eq. (1)]. We approximate the contact force using a simple spring-model, in which the force acting on either object is proportional to the displacement of the surface at the point of contact. This yields the force between two objects as a half-wavelength of a sinusoid

$$f(t) = \begin{cases} \sin\left(\sqrt{\frac{k}{m}}t\right) & \forall \;\; 0 < t < \frac{\pi m}{k} \\ 0 & \text{otherwise} \end{cases} \quad , \tag{5}$$

where $v$ is the velocity at impact, $m$ the mass of the pellet and $k$ a spring constant determined by the materials of the board and ball. Note that as the mass tends to zero, the time of contact between the two materials tends to zero and the contact force tends towards a Dirac-delta function. This observation partly justifies the use of small pellet impact recordings to approximate the object impulse response.

To synthesize impact sounds, we convolve a synthesized IR from Eq. (4) with the contact force described in Eq. (5). All that needs to be encoded at the time of impact are labels of object mass, velocity, and material labels, which determine both the spring constants and the distributions from which modes are sampled. Except for parameters of the mode distributions, these features are already included in physics engines.

## 3. PERCEPTION OF SYNTHETIC IMPACTS

To assess our impact synthesis model we played both recorded and synthesized sounds to listeners and asked them to judge: (1) realism; (2) material; and (3) mass of the colliding objects. All perceptual experiments were conducted over Amazon's Mechanical Turk platform. A standardized test was used to ensure participants were wearing headphones [40].

### 3.1. Experiment 1. Realism of synthetic impact sounds

We first sought to test whether our synthetic sounds were compelling renditions of real-world impacts. If our synthesis method neglected sound features to which the brain is sensitive, the synthetic sounds should be recognizable as fake.

Participants were presented with a pair of impact sounds and identified which was the real recording. In all trials, one sound was a real-world recording of a ball dropped on a resonant object, and one a synthetic impact generated via our model or a model that was 'lesioned' in some way, by omitting the transient component of the IR, or by omitting the modes from the IR. The conditions of the experiment were (1) full synthetic model; (2) Modes only, without transient; (3) Transient only, without modes; (4) Time-reversed synthetics. The sound in the final condition were clearly synthetic, which serves to ensure task comprehension.

The results (Fig. 2) show that listeners could not distinguish sounds from either the full or lesioned models from real-world recordings, demonstrating that our method of impact sound synthesis yields plausible sounds. The chance performance for the lesioned models presumably reflects the fact that the resulting sounds remained realistic even though the lesion altered the quality of the sounds. As participants were good at identifying the Time-Reversed sounds it is clear they understood the task. Poor performance in the other conditions thus reflects the success of the synthesis.

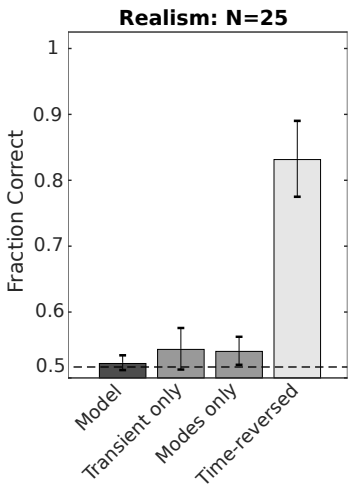


Figure 2: *Discrimination of real vs. synthetic impact sounds (Exp 1). Dashed line denotes chance performance.*

### 3.2. Experiment 2. Perception of material

Having demonstrated that our synthetic impact sounds are realistic, we sought to test whether they convey appropriate physical parameters to listeners. We first tested whether listeners can recognize the material of a struck resonant object.

Participants heard a single impact sound and were asked to identify the material of the struck object from one of four possible categories: metal, ceramic, wood or cardboard. Participants were told that the striking mallet was effectively noiseless and that many different objects of each material class were used, of a range of

different sizes, shapes and sub-material (i.e. metal contained steel, tin, aluminium etc.; wood contained poplar, pine, oak etc.)

With real-world recordings, participants were excellent at distinguishing hard materials (metal or ceramic) from soft materials (wood or cardboard) but made errors within the hard or soft categories (Fig. 3). This result is consistent with prior studies [13]. Sounds from our synthesis model - both with and without the transient - yielded a similar pattern of success and failures. Without modes, or with shortened modes, human judgments were strongly biased towards to softer materials. With lengthened modes, judgments were biased towards harder materials, particularly metal. This demonstrates that our model - particularly the mode statistics - have captured the acoustical features that humans use to judge material classes from impact sounds. The correlation of the confusion matrices for the full model and recorded sounds was 0.72.

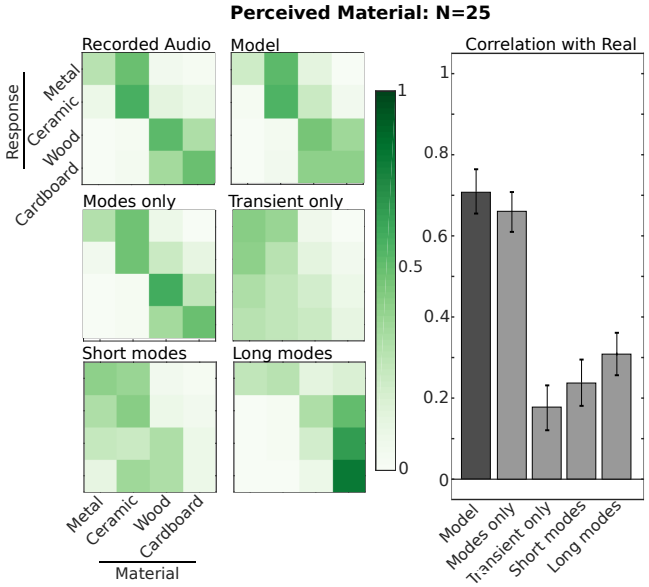


Figure 3: *Material discrimination from synthetic impact sounds (Exp 2). Left: Confusion matrices of the presented material and participant responses. Right: Correlation of the confusion matrices of various synthetic sounds with that of the recorded impacts.*

### 3.3. Experiment 3. Perception of mass

We next sought to test whether our synthetic sounds convey the mass of the striking mallet to listeners. Participants heard two impact sounds, one of a small wooden pellet (0.7 g) dropped onto an object, and one of a larger wooden ball (7.6 g) dropped onto the same object. Participants were asked to identify which of the two balls was heavier. To generate synthetic sounds the synthetic IRs were convolved with two different contact forces to emulate different ball masses, as shown in Eq. (5). The impact levels were not normalized but retained the relative variation in power level induced by the difference in impact force (i.e. the coefficient in Eq. (5) and the amplitude of the IR). All recordings and simulations were made with balls dropped from the same height (8 cm), but participants were not explicitly told this.

Since we do not know $k$, the spring constant, we cannot compute the contact force [Eq. (5)]. Instead we estimate $k$ from the recorded impact sounds. Since both balls are the same material, we assume $k_{\text{large}} = k_{\text{small}}$, which means the ratio between the contact times for the two balls is $m_{\text{large}}/m_{\text{small}}$. We set the contact time of the larger ball to be 10.9 times that of the smaller ball. We then iteratively adjusted the contact time of the smaller ball, until it produced a match between the average spectral centroid of the synthetic sounds and of the corresponding impact recordings.

The results (Fig. 4) show that humans perform very well at this task, both with real-world recordings and with synthetic sounds. This demonstrates that humans are sensitive to the filtering effect described by the contact force and can use this acoustic information to estimate the mass of the striking mallet. Participants showed a small performance decrement in the conditions where modes were shortened or excluded altogether, suggesting that humans are using modes, in addition to the sound level and spectral centroid, to estimate mass. The results suggest that our synthetic sounds convey mass as well as real-world recordings.

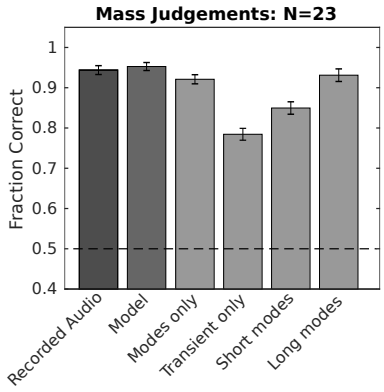


Figure 4: *Mass discrimination with real and synthetic impact sounds (Exp 3).*

## 4. SUSTAINED CONTACTS

To test the generalizability of our impulse response distributions, we next consider sustained contacts such as made by two objects scraping across each other. Similar to [32], we again use the source-filter model of Eq. (1) but both the force and object IRs are more complicated than for impact sounds. The contact force $f(t)$ is generated by a series of small collisions as the scraper moves across the surface of the scraped object, and is thus a function of the downward force applied to the scraper, the surface texture depth, and the scraper speed (Fig. 1, bottom). The object IR changes with scraper position $x(t)$, and thus, as the scraper moves across the surface, the IR becomes a time-varying function $h_{\text{surface}}(x(t))$. We describe these models of force and IR in more detail below. Despite the simplicity of this model, our results suggest that it yields plausible scraping sounds which convey motion of the scraper.

### 4.1. Contact force for sustained contacts

To model the force between scraper and surface we start with several simplifying assumptions: that the external force applied to the scraper $F_{\text{p}}$ is constant and applied vertically downwards, and that the probe follows the surface exactly without any slip or bounce,

such that the probe height $z(t)$ at time $t$, is given by the surface elevation $S(x)$ at the probe location $x$. For now we consider a transect across the surface so $x$ is a one-dimensional variable, though the following analysis applies easily to a 2D treatment.

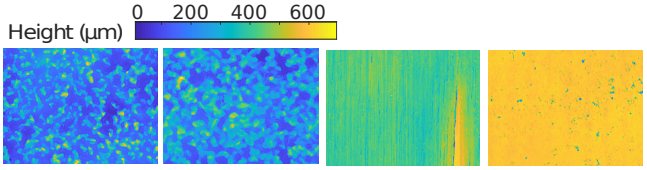


Figure 5: *Everyday textures measured with the confocal microscope. Surface area is 7.3 mm by 10 mm. From-left: 100 grit sandpaper; 60 grit sandpaper; wood; vinyl tile.*

We first consider the vertical component of the force. Under our assumptions, the change in vertical force applied to the surface can be derived from the vertical acceleration of the probe, which, as the probe follows the surface, is given by

$$\begin{aligned} f_{\mathrm{v}}(t) &= m_{\mathrm{p}}\ddot{z} \\ &= m_{\mathrm{p}}\frac{\partial^2 S}{\partial x^2}\,|v(t)|^2 \quad , \end{aligned} \tag{6}$$

where $m_{\mathrm{p}}$ is the mass of the probe and $v(t)$ is the horizontal velocity of the probe.

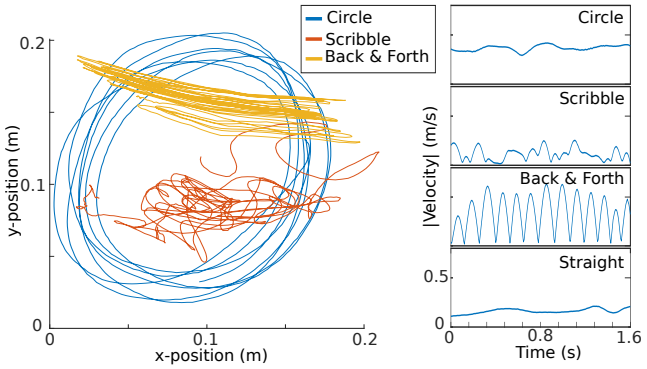


Figure 6: *Scraping motions. Left: Measured position traces of scraper over surface, for three different types of motion. Right: Absolute velocity measurements.*

We next consider the frictional force tangential to the surface. We model this as proportional to the probe speed raised to the power of an exponential factor $\gamma$, giving

$$f_{\mathrm{h}}(t) \propto \left(\left|v(t)\frac{\partial S}{\partial x}\right|\right)^{\gamma} \quad , \tag{7}$$

where the partial derivative with respect to x accounts for the difference between the speed of scraping across the surface, which is the important factor, and the horizontal speed $|v(t)|$. The total force imparted onto the object is then given by

$$\begin{aligned} f(t) &= f_{\mathrm{v}} + f_{\mathrm{h}} \\ &= m_{\mathrm{p}}\frac{\partial^2 S}{\partial x^2}v(t)^2 + A\left(\left|v(t)\frac{\partial S}{\partial x}\right|\right)^{\gamma} \quad , \end{aligned} \tag{8}$$

where $A$ and $\gamma$ are unknown constants which titrate the importance of shear friction versus vertical forcing. We explore the role of these factors by listening to synthetic scrape sounds from a range of values. We have neglected the constant downward force term $F_{\mathrm{p}}$ which, though present, does not create any sound.

To obtain $S$, we measured the surface texture of several real objects using a scanning confocal microscope (Keyence VK-X260K). In these experiments, we used a micro-scale depth map of a small section of a wood block (Fig. 5). These are relatively small matrices (1600 pixels by 2300 pixels), which render the surface with horizontal resolution of 5.6 $\mu$m and vertical resolution of 0.1 nm. Based on perceptual results concerning auditory texture perception, we expect that the perceptually important properties of such textures are statistical [26]. Therefore, to define $S$, we use one-dimensional quilting to generate a texture from a measured depth map [10], sampling a series of single rows and concatenating them. In future work, we plan to synthesize these surfaces statistically from coarse-level variables, in the same spirit as our distribution over impulse responses.

In addition to a depth map, the synthetic scraping force requires ecologically plausible velocity profiles of scraping motions. To probe the mechanics of typical human scraping movements, we measured the velocity and position profiles of several scraping movements using an optical tracking system (OptiTrack V120:Trio; Fig. 6). We use these recorded trajectories in the reported synthesis. However, in informal experiments, we found that the quality of sound synthesis was not heavily dependent on a precise match to the recorded data. Future work will include simple statistical models of these trajectories.

### 4.2. Variation of IRs over contact location

Object IRs depend upon the location being struck, and thus to simulate scraping we model this variation of modal properties with probe location. To informally assess the variability of mode properties as a function of impact location, we compared impact recordings we had made with different strike locations and found the variation in mode properties to be moderate. To emulate such changes with synthetic IRs, we synthesized a single canonical IR from our model [Eq. (4)], with properties $(\mathbf{a}_o, \mathbf{b}, \omega)$, and simulated a number of location specific IRs by adding some noise to the mode powers

$$(\vec{a}, \vec{b}, \vec{\omega}) = (\vec{a_{\mathbf{o}}} + \vec{\epsilon}, \vec{b}, \vec{\omega}) \quad , \tag{9}$$

where $\vec{a_{\mathbf{o}}}$ is the original vector of mode powers sampled from our model and $\vec{\epsilon}$ is a Gaussian noise vector sampled with zero-mean and a standard-deviation set to 20% the mean mode onset power. This gives a set of IRs with similar but varying modes, which crudely emulate an object of arbitrary shape struck in various locations.

We assign these sampled IRs to points along a motion trajectory, and interpolated between them in waveform space to give a smoothly varying surface IR, $h_{\mathrm{surface}}(x(t))$. When the scraper was at a position between the defined centerpoints, the impulse response was a linear combination of the impulse responses with weights proportional to the relative distances from the scraper to the centerpoints. We ignore the contribution of the scraper to the impulse response, assuming that it is damped by the hand in which it is held.

## 5. PERCEPTION OF SYNTHETIC SCRAPING

To assess the efficacy of our scraping synthesis model, we played both recorded and synthesized sounds to listeners and asked them to judge: (1) realism; and (2) the shape of the scraper's position trajectory. As in section 3, all experiments were conducted online using Amazon's Mechanical Turk platform, and a standardized test was used to ensure participants were wearing headphones [40]. In each experiment, in addition to testing lesioned forms of our own synthesis model, we compare our model to the one other scraping synthesis method that we are aware has been tested psychophysically [35]. Thoret et al. generated low-pass filtered white noise whose amplitude and filter cutoff increased with increasing velocity, and showed that several motion trajectories could be accurately judged from the resulting sounds.

### 5.1. Experiment 5. Realism of synthetic scraping sounds

Participants were played a pair of scraping sounds and asked to identify which was the real recording. In all trials, one sound was a real-word recording of chopstick scraping a board, and one a synthetic scrape generated via our model or a lesioned version thereof. The synthetic conditions of the experiment were generated via (1) the full model, using measurement-based surface textures and varied IRs; (2) measured depth map and just a single IR; (3) pink noise depth map and varied IRs; (4) white noise with varied filter cutoff from [35]; and (5) velocity-gated white noise, which is silent when the chopstick moves more slowly than a threshold, but otherwise constant. Condition (5) is clearly synthetic and serves to ensure the participants understand the task.

The results (Fig. 7) show that the full synthesis model, though not perfectly realistic, frequently fools listeners. However, using a time-varying impulse response does not improve realism over filtering with a single synthetic impulse response. A synthetic noise depth map also produced comparably realistic sounds. Our sounds were less obviously synthetic than those of [35], but one caveat is that the comparison recordings were produced by a narrow scraping probe. We suspect that condition (4), with its flat broadband spectrum, may be more appropriate for modeling scrapes produced by heavier objects with large contact surface area (e.g. pushing a heavy box over tile). The gated white-noise is easily recognized as synthetic by the participants, demonstrating that they understood the task.

### 5.2. Experiment 6. Perception of motion

Participants were presented with a single scraping sound and asked to choose the scraping trajectory from four choices: "circular", "back-and-forth", "scribble", or "straight". Participants heard both real-world recordings and synthetic sounds derived from a real-world motion. The motion trajectories used to generate synthetic scrapes were matched in speed to the scrapers used to make the real-world recordings.

As shown in Fig. 8, motion judgments for synthetic scrapes were similar to those for real-world scrape recordings. In both cases participants were correct most of the time, but misjudged "straight" motions to be "circular", both of which have velocity profiles without zero points. When judging either "back-and-forth" or "scribble" sounds, the full model and its lesioned variants led to more "scribble" judgments. This result could reflect the greater scattering of contact position around the surface in scribbling compared to other motions. Although we attempted to simulate this positional change with changing IRs, the full model and the constant IR model were comparable for both realism and motion, suggesting that we did not successfully capture this informative spatial variation.

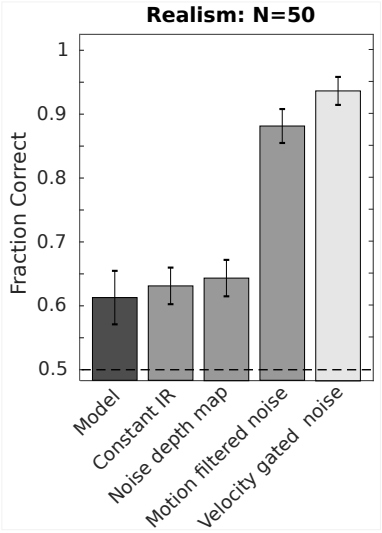


Figure 7: *Discrimination of real vs. synthetic impact sounds scraping sounds (Exp 5). Dashed line indicates chance performance.*

## 6. DISCUSSION

Our synthesis model is fast because it only models the effects of a small number of physical variables (material, mass, velocity etc.). It is evident from daily life that humans can infer more than just the variables we have described from contact sounds. Impact sounds contain cues to shape, size, and hollowness, as well as to the environmental reverberation [36]. Some physical variables explored in our impact model can also be conveyed by frictional sounds (e.g. material) but this remains to be explored in future work. Furthermore, friction sounds are not limited to scraping, but rather include other interactions such as rubbing, brushing and sliding. Future investigations into how these interactions produce sound, and into human sensitivity to their properties, will hopefully suggest extensions to a better and more nuanced synthesis algorithm.

The current version of our synthesis model requires some physical measurements of real-world objects: statistical distributions of object IRs conditioned upon material parameters; and surface structures. In future we hope to be able to synthesize these intermediate representations from physical variables. Our impact experiments with altered IRs demonstrated that lengthening or shortening the resonant modes caused listeners to rate the synthetic materials as"harder" or "softer" materials, consistent with physical models [13, 17, 21], but did not diminish their realism. This suggests that we should be able to synthesize IRs for novel objects without having to measure them first, permitting sound synthesis for a much larger range of objects. Similar generalizations should be possible for the forcing functions used to generate scraping sounds. As with perception of acoustic textures [25], it is likely that humans are insensitive to the fine-grained temporal details of the contact force we use to synthesize scrapes. Presumably we can synthesize such a contact force directly from a texture model [26],

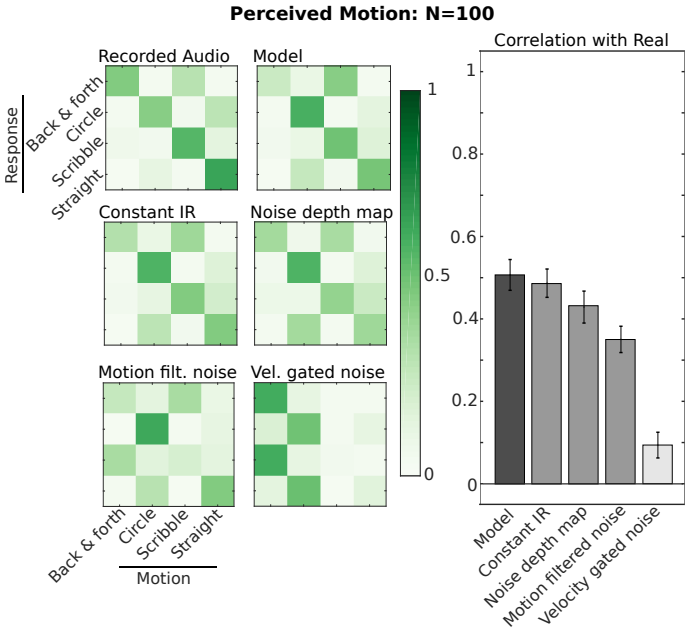


Figure 8: *Motion discrimination from synthetic scrape sounds (Exp 6). (Left) Confusion matrices of presented motion pattern and the human responses. (Right) Correlations of the confusion matrices of synthetic sounds with the correlation matrix of recorded sounds.*

enabling sound synthesis for a wider and more diverse range of objects without costly and time-consuming measurements.

Our impulse response model, while derived from statistics of impact sounds, can successfully contribute to the synthesis of relatively realistic scraping sounds. However, it appears that this model does not accurately capture the spatial covariance between impulse responses over a surface. Our full model and lesioned model with a single IR perform equally well, and neither are yet on par with real recordings, both in terms of realism and in the conveyed motion (Fig. 7, Fig. 8). Future investigations will include measurements and modelling of this variation in impulse responses based on position, as well as comparing modes measured from scraping sounds with those from impacts. The other component of the scraping synthesis is an excitation force based on quilted textures of measured depth maps. Several authors have treated scraping as a noisy source paired with a modal filter. Some model the friction force as $1/f^{\beta}$ noise [9, 32, 39], while others use a statistical model of densely-packed impact events [18]. In the experiments explored here, the utilization of real-world measurements did not improve realism or motion inference. However, it remains possible that constraining a more sophisticated model of surface texture with these measurements could be useful, particularly in judgments of material and surface roughness.

The model we have presented is similar in some respects to that of Conan et al. [8], who used statistics of contact forces to synthesize rolling sounds. We also utilize a statistical approach, but model the sounds of impacts and scraping, using statistics of the resonant modes of objects. We also found that we could use a linear model for contact forces. By contrast, Conan et al. found that a non-linearity in impact force (namely that the duration of impact should change with impact force) was required to induce realistic rolling sounds. In the future, we plan to investigate whether there are perceptual benefits to sound synthesis with more realistic impact forces.

## 7. CONCLUSION

We have presented a fast and efficient method for synthesis of contact sounds - inspired by both physics and perception. The method generates object IRs by sampling resonant modes from distributions fitted to empirical measurements from example impact sounds. The method then convolves the IRs with contact force simulated with a simple physics model of either impacts or sustained scrapes. Despite the simplicity of the model, perceptual listening tasks demonstrate that the synthetic sounds are realistic and convey basic physical information as well as recorded sounds. These results suggest that our model has captured many of the acoustic features that matter for perception of physical contact sounds, despite neglecting a great deal of physical information about the sound sources.

## 8. REFERENCES

[1] Mitsuko Aramaki, Mireille Besson, Richard Kronland-Martinet, and Sølvi Ystad. Controlling the perceived material in an impact sound synthesizer. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):301–314, 2010.

[2] Federico Avanzini and Davide Rocchesso. Controlling material properties in physical models of sounding objects. In *ICMC*, 2001.

[3] Stefan D Bilbao. *Numerical sound synthesis*. Wiley Online Library, 2009.

[4] Claude Cadoz. *Synthèse sonore par simulation de mécanismes vibratoires. Applications aux sons musicaux*. PhD thesis, Institut national polytechnique de Grenoble, 1979.

[5] Claude Cadoz, Annie Luciani, and Jean-Loup Florens. Responsive input devices and sound synthesis by stimulation of instrumental mechanisms: The cordis system. *Computer music journal*, 8(3):60–73, 1984.

[6] Claude Cadoz, Annie Luciani, and Jean Loup Florens. Cordis-anima: a modeling and simulation system for sound and image synthesis: the general formalism. *Computer music journal*, 17(1):19–29, 1993.

[7] Claudia Carello, Krista L Anderson, and Andrew J Kunkler-Peck. Perception of object length by sound. *Psychological science*, 9(3):211–214, 1998.

[8] Simon Conan, Olivier Derrien, Mitsuko Aramaki, Sølvi Ystad, and Richard Kronland-Martinet. A synthesis model with intuitive control capabilities for rolling sounds. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(8):1260–1273, 2014.

[9] Simon Conan, Etienne Thoret, Mitsuko Aramaki, Olivier Derrien, Charles Gondre, Solvi Ystad, and Richard Kronland-Martinet. Intuitive synthesizer of continuous-interaction sounds: Rubbing, scratching, and rolling. *Computer Music Journal*, 38(4):24–37, 2014.

[10] Alexis A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. *SIGGRAPH: Computer Graphics*, 2001.

[11] Neville H Fletcher and Thomas D Rossing. *The physics of musical instruments*. Springer Science & Business Media, 2012.

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

[13] Bruno L Giordano and Stephen McAdams. Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *The Journal of the Acoustical Society of America*, 119(2):1171–1181, 2006.

[14] Massimo Grassi. Do we hear size or sound? balls dropped on plates. *Perception & psychophysics*, 67(2):274–284, 2005.

[15] Hermann LF Helmholtz and Alexander J Ellis. On the sensation of sound in general. 1875.

[16] Doug L James, Jernej Barbič, and Dinesh K Pai. Precomputed acoustic transfer: output-sensitive, accurate sound generation for geometrically complex vibration sources. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 987–995. ACM, 2006.

[17] Roberta L Klatzky, Dinesh K Pai, and Eric P Krotkov. Perception of material from contact sounds. *Presence: Teleoperators & Virtual Environments*, 9(4):399–410, 2000.

[18] Mathieu Lagrange, Gary Scavone, and Philippe Depalle. Analysis/synthesis of sounds generated by sustained contact between rigid objects. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):509–518, 2010.

[19] Guillaume Lemaitre and Laurie M Heller. Auditory perception of material is fragile while action is strikingly robust. *The Journal of the Acoustical Society of America*, 131(2):1337–1348, 2012.

[20] D Brandon Lloyd, Nikunj Raghuvanshi, and Naga K Govindaraju. Sound synthesis for impact sounds in video games. In *Symposium on Interactive 3D Graphics and Games*, pages PAGE–7. ACM, 2011.

[21] Robert A Lutfi. Human sound source identification. In *Auditory perception of sound sources*, pages 13–42. Springer, 2008.

[22] Robert A Lutfi and Christophe NJ Stoelinga. Sensory constraints on auditory identification of the material and geometric properties of struck bars. *The Journal of the Acoustical Society of America*, 127(1):350–360, 2010.

[23] Dinesh Manocha and Ming C Lin. Interactive sound rendering. In *2009 11th IEEE International Conference on Computer-Aided Design and Computer Graphics*, pages 19–26. IEEE, 2009.

[24] Stephen McAdams, Vincent Roussarie, Antoine Chaigne, and Bruno L Giordano. The psychomechanics of simulated sound sources: Material properties of impacted thin plates. *The Journal of the Acoustical Society of America*, 128(3):1401–1413, 2010.

[25] Josh H McDermott, Michael Schemitsch, and Eero P Simoncelli. Summary statistics in auditory perception. *Nature neuroscience*, 16(4):493, 2013.

[26] Josh H McDermott and Eero P Simoncelli. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–940, 2011.

[27] Philip McCord Morse and K Uno Ingard. *Theoretical acoustics*. Princeton university press, 1986.

[28] James F O'Brien, Chen Shen, and Christine M Gatchalian. Synthesizing sounds from rigid-body simulations. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 175–181. ACM, 2002.

[29] Laurent Pruvost, Bertrand Scherrer, Mitsuko Aramaki, Sølvi Ystad, and Richard Kronland-Martinet. Perception-based interactive sound synthesis of morphing solids' interactions. In *SIGGRAPH Asia 2015 Technical Briefs*, page 17. ACM, 2015.

[30] Nikunj Raghuvanshi and Ming C Lin. Interactive sound synthesis for large scale environments. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, pages 101–108. ACM, 2006.

[31] John William Strutt Baron Rayleigh. *The theory of sound*, volume 1. Macmillan, 1896.

[32] Zhimin Ren, Yeh Hengchin, and Ming C. Lin. Synthesizing contact sounds between textured models. University of North Carolina at Chapel Hill, 2010.

[33] Zhimin Ren, Hengchin Yeh, and Ming C Lin. Example-guided physically based modal sound synthesis. *ACM Transactions on Graphics (TOG)*, 32(1):1, 2013.

[34] Davide Rocchesso and Federico Fontana. *The sounding object*. Mondo estremo, 2003.

[35] Etienne Thoret, Mitsuko Aramaki, Richard Kronland-Martinet, Jean-Luck Velay, and Solvi. Ystad. From sound to shape: auditory perception of drawing movements. *Journal of Experimental Psychology: Human Percept ion and Performance*, 40(3):983–994, 2014.

[36] James Traer and Josh H McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016.

[37] Simon Tucker and Guy J Brown. Investigating the perception of the size, shape and material of damped and free vibrating plates. *University of Sheffield, Department of Computer Science Technical Report CS-02-10*, 2002.

[38] Kees van de Doel and Dinesh K Pai. Synthesis of shape dependent sounds with physical modeling. Georgia Institute of Technology, 1996.

[39] Kees van den Doel, Paul G. Kry, and Dinesh K. Pai. Foleyautomatic: Physically-based sound effects for interactive simulation and animation. University of British Columbia, 1996.

[40] Kevin JP Woods, Max H Siegel, James Traer, and Josh H McDermott. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, 79(7):2064–2072, 2017.

[41] Changxi Zheng and Doug L James. Toward high-quality modal contact sound. In *ACM Transactions on Graphics (TOG)*, volume 30, page 38. ACM, 2011.