PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

# Intuitive knowledge of object acoustics enables perceptual separation of physical variables from impact sounds

Vinayak Agarwal[a,b,c,1], James Traer[a,c,d,e,1], Jeremy Schwartz[a], and Josh H. McDermott[a,c,f,2]

Affiliations are included on p. 12.

Edited by Laurie M. Heller, Carnegie Mellon University, Pittsburgh, PA; received February 19, 2026; accepted October 2, 2026 by Editorial Board Member Roberta L. Klatzky

**Upon hearing objects collide, humans can estimate physical attributes such as material and mass. Although the physics of sound generation is well established, the inverse problem that listeners solve—of inferring physical parameters from sound—remains poorly understood. Classical accounts posit the use of acoustic cues that correlate with physical variables but do not explain how humans might distinguish multiple concurrent physical causes. To study this problem, we measured impulse responses from hundreds of everyday objects, extracted their resonances, and used statistics of the resonances to build a probabilistic generative model of impact sounds. We then used the model to synthesize and manipulate stimuli for experiments measuring judgments of physical variables. Humans accurately judged object properties from collision sounds. However, when both of the colliding objects varied, performance was impaired if the distribution of object resonances deviated from those measured in real-world objects. The results suggest that humans have internalized the statistical regularities of real-world object acoustics, and use this knowledge to separate the acoustic contributions of the distinct objects in the world that give rise to sounds.**

auditory scene analysis | intuitive physics | ecological acoustics

## Significance

We investigate how humans infer physical properties from sound. First, we present a large-scale dataset of acoustic measurements from over 400 everyday objects, enabling estimates of the distribution of object resonances and probabilistic models that generate realistic impact sounds from physical variables. Second, we show that listeners can disentangle multiple physical variables from a single sound–for example, judging which of two balls is heavier even when dropped on different surfaces. Third, we show that this ability depends on implicit knowledge of natural acoustic regularities and breaks down when object resonances are drawn from unnatural distributions. These findings suggest that humans have internalized real-world acoustic statistics and use them to infer the physical causes of sounds.

When two objects collide, they create a sound, from which humans can judge their physical properties. Consider for example, a plate set on a table. Everyday experience suggests that a listener can identify aspects of object shape, motion, and material (1).

The creation of sound in such settings is well studied (2). Objects are known to vibrate most strongly at certain frequencies known as "resonant modes" (3, 4), the frequencies, powers and decay rates of which are determined by the object shape and material (5). High-resolution acoustical models now allow synthesis of realistic impact sounds for a wide range of complex objects (6–8), given precise specification of the object geometry and material.

By contrast, the inverse problem solved by perception–of inferring physical properties from sound–remains poorly understood. Prior work shows that humans can estimate some physical parameters from impact sounds: material (9–15), size (16, 17), force of impact (18), and aspects of an object's shape (19, 20). However, the basis of these judgments is unclear.

The most common explanation of these abilities is that humans have learned to associate particular acoustic cues with physical variables. For instance, the hardness of a struck object has often been linked to the ratio of the decay rate of sound to its frequency (10). Similarly, the mass of a dropped object has been linked to spectral centroid (the "center of mass" of the spectrum) and overall sound level (16, 17, 20).

One challenge for cue-based explanations is that sound is always produced by multiple physical variables. Consider an object dropped on a surface (Fig. 1*A*). The resulting sound is a function of the resonant properties of the surface (3, 4), but also of the force created at impact (18, 21), which depends on the height from which the object was dropped, its mass (16), and the object and surface stiffnesses. As a result, it is not obvious that any single cue can be diagnostic of any single physical variable, because the cue is likely to be influenced by other concurrent physical variables.

An alternative perspective is that estimating any individual physical factor requires "separating" its causal contribution from that of the other physical factors that determine a sound. Under this view, the problem of physical inference is analogous to the "cocktail party problem" (22) in which a listener must separate a single voice from a mixture of voices, or to the problem of separating environmental reverberation from the sound sources in a room (23), each of which involve multiple causal variables. However, much previous research on auditory perception of physical variables has varied only one physical variable within an experiment, leaving it unclear a) whether humans can judge physical variables

Author contributions: V.A., J.T., and J.H.M. designed research; V.A., J.T., J.S., and J.H.M. performed research; V.A. analyzed data; and V.A., J.T., and J.H.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. L.M.H. is a guest editor invited by the Editorial Board.

Copyright © 2026 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY).

[1]V.A. and J.T. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: jhm@mit.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2605876123/-/DCSupplemental.

Published November 6, 2026.

Downloaded from https://www.pnas.org by 86.130.54.74 on November 6, 2026 from IP address 86.130.54.74.

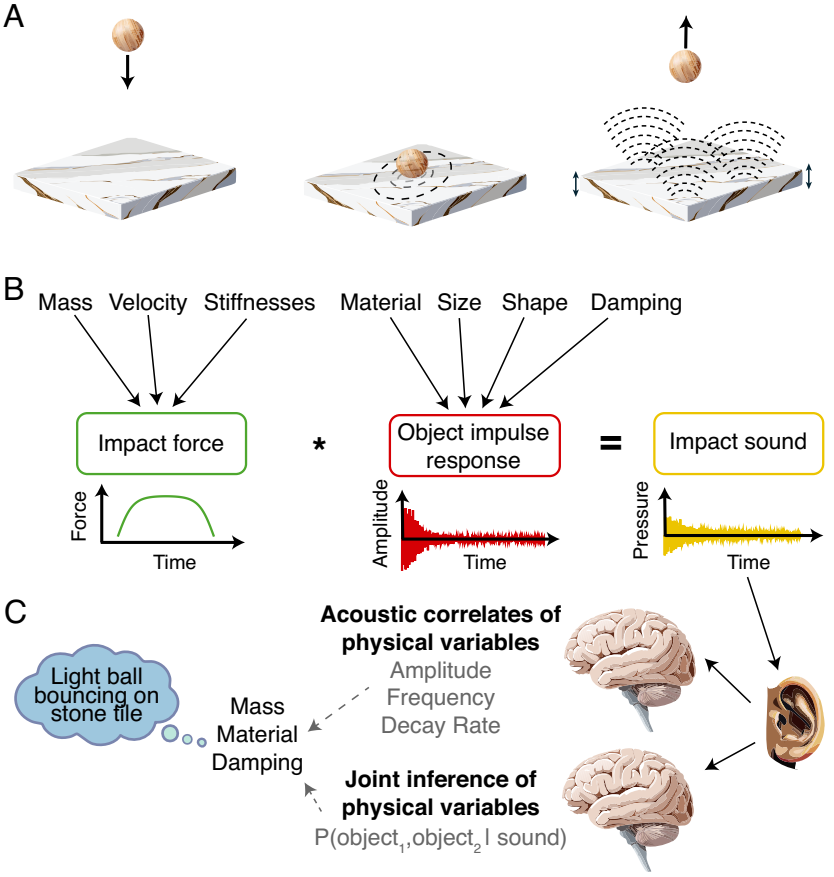



**Fig. 1.** Generation and perception of impact sounds. (*A*) Impact sound produced by a ball striking a surface. When the ball contacts the surface, a momentary force is created, inducing vibrations in the surface that resonate within it and then radiate into the surrounding medium, causing the sound that we hear. Vibrations are also induced in the ball, but if the ball is small these may produce negligible sound in comparison to the struck surface. (*B*) Linear systems model of impact sound generation, with dependencies between physical variables and signals underlying sound generation. An impact sound can be approximated as the convolution of a time-varying impact force and the impulse response of the vibrating surface (* denotes convolution). The force is a function of the ball mass and velocity as well as the stiffnesses of the ball and surface. The impulse response of the surface is a function of its properties, including material, size, shape, and damping. The presence of these multiple causal influences on the resulting sound renders the problem of inferring the properties of either object ill-posed. (*C*) Alternative explanations of human physical inferences. A common proposal (*top*) is that humans have learned acoustic correlates of physical variables, and use these to estimate these variables. An alternative (*bottom*) is that humans jointly infer (i.e., perceptually "separate") the properties of the underlying objects, using some type of knowledge of how sound is generated from physical variables.

from sound amid multiple physical causes, and b) how humans might solve the underlying inference problem.

Inference problems involving multiple underlying causes are usually ill-posed (lacking a unique solution absent other constraints). Ill-posed problems in perception are typically solved using implicit knowledge of the regularities of the underlying causal factors (24), as these regularities narrow the space of possible solutions. We hypothesized that humans have internalized physics-induced regularities in how sound is generated, and use this intuitive knowledge when estimating the physical causes of a sound. To test this hypothesis, we first conducted a large-scale acoustic study of everyday objects to characterize physics-induced regularities in object sounds. We then asked humans to make physical inferences from the sound of a ball striking a surface, manipulating the statistics of object resonances to test reliance on implicit knowledge of object acoustic regularities.

## Results

**Generation of Impact Sounds.** Consider a ball dropped on a surface (Fig. 1*A*). Force is created when the ball impacts the surface and then bounces away. In many situations, the objects are well approximated as linear systems (4). Their vibrational response can thus be modeled by the convolution of their impulse responses and a time-varying force. An impulse response is defined by the point of contact (where the force is applied) and the point of sound measurement. For now, consider impulse responses defined with respect to a point of measurement in the surrounding medium (as if a microphone was near the objects). In the simplest case with a single impact (no subsequent bounces), the sound can be modeled as:

$$s(t) = f(t) * \left[h_{\text{surface}}(t) + h_{\text{ball}}(t)\right],$$

where $s$ is the impact sound, $f$ is the contact force, $h_{\text{surface}}$ is the impulse response of the impacted surface, and $h_{\text{ball}}$ is the impulse response of the ball. If the ball is small and dense, it will radiate much less sound than the surface (16, 17, 25), and the contribution of its impulse response can be neglected:

$$s(t) = f(t) * h_{\text{surface}}(t).$$

In real-world settings both $f$ and $h_{\text{surface}}$ depend on underlying physical variables (Fig. 1*B*).

We adopt this problem setting to study the perceptual inference of physical variables from sound, as it retains the ill-posedness of the general problem (because the force depends on both objects and their momentum at impact), while reducing the number of factors to consider and manipulate. We seek to distinguish two classes of hypotheses (Fig. 1*C*). One holds that humans derive acoustic cues from an impact sound and use those to judge physical variables. The other holds that humans use knowledge of the sound generating process (specifically, the distribution of impulse responses) to jointly infer physical variables. The second class of hypothesis does not preclude the use of acoustic cues, but posits that any such cues are interpreted in light of the distribution of impulse responses. Although inclusive of inference in a rich generative model of physics, this hypothesis class also encompasses simpler accounts that use knowledge of physics-related acoustic regularities to estimate likely combinations of multiple physical variables.

To distinguish these two hypothesis classes, we asked whether the ability to infer one physical variable (e.g., ball mass) is influenced by the ability to infer a concurrent variable (e.g., surface material). We manipulated the ability to infer the concurrent variable by altering the extent to which one object's impulse response conformed to the real-world distribution (of impulse responses). The rationale was that this conformity should be necessary to infer

Downloaded from https://www.pnas.org by 86.130.54.74 on November 6, 2026 from IP address 86.130.54.74.

the concurrent variable given the ill-posedness of the problem. In the sections that follow we characterize the distribution of everyday object impulse responses using large-scale acoustic measurements, use the measurements to estimate a distribution over object impulse responses, and build a generative model of impact sounds. We then use the generative model to manipulate impulse response statistics in perceptual experiments.

**Object Impulse Response Measurement.** Impulse responses were measured from everyday objects by broadcasting pseudorandom noise into an object and rerecording the noise from the object's surface, analogous to methods often used to measure the impulse responses of reverberant spaces (23). We measured impulse responses from the object surface in order to make measurements in everyday environments (for which external microphone recordings would be contaminated by environmental reverberation). Contact speakers and microphones attached to the object's surface were used to transmit vibrations into the object and record the result, from which the impulse response could be estimated (Fig. 2*A*; see *Materials and Methods*).

To select objects that span the range encountered in daily life, we used an online survey. Participants reported the objects for which they had recently heard collision sounds, with separate prompts for each of five size categories ranging from less than 10 cm to greater than 3 m in linear dimension. We combined their responses into a target list of objects (*SI Appendix*, Table S1) and recorded as many of these commonly identified objects as was practical (see *Materials and Methods* and *SI Appendix*, Table S2). The resulting set of objects spanned a wide range of material properties, sizes, and shapes (Fig. 2*B*). To our knowledge these measurements constitute the first large-scale characterization of everyday object acoustics.

**Real-World Object Measurements Reveal Stereotyped Form of Object Impulse Responses.** There was considerable consistency in the form of the impulse responses. The example shown in Fig. 2*C*

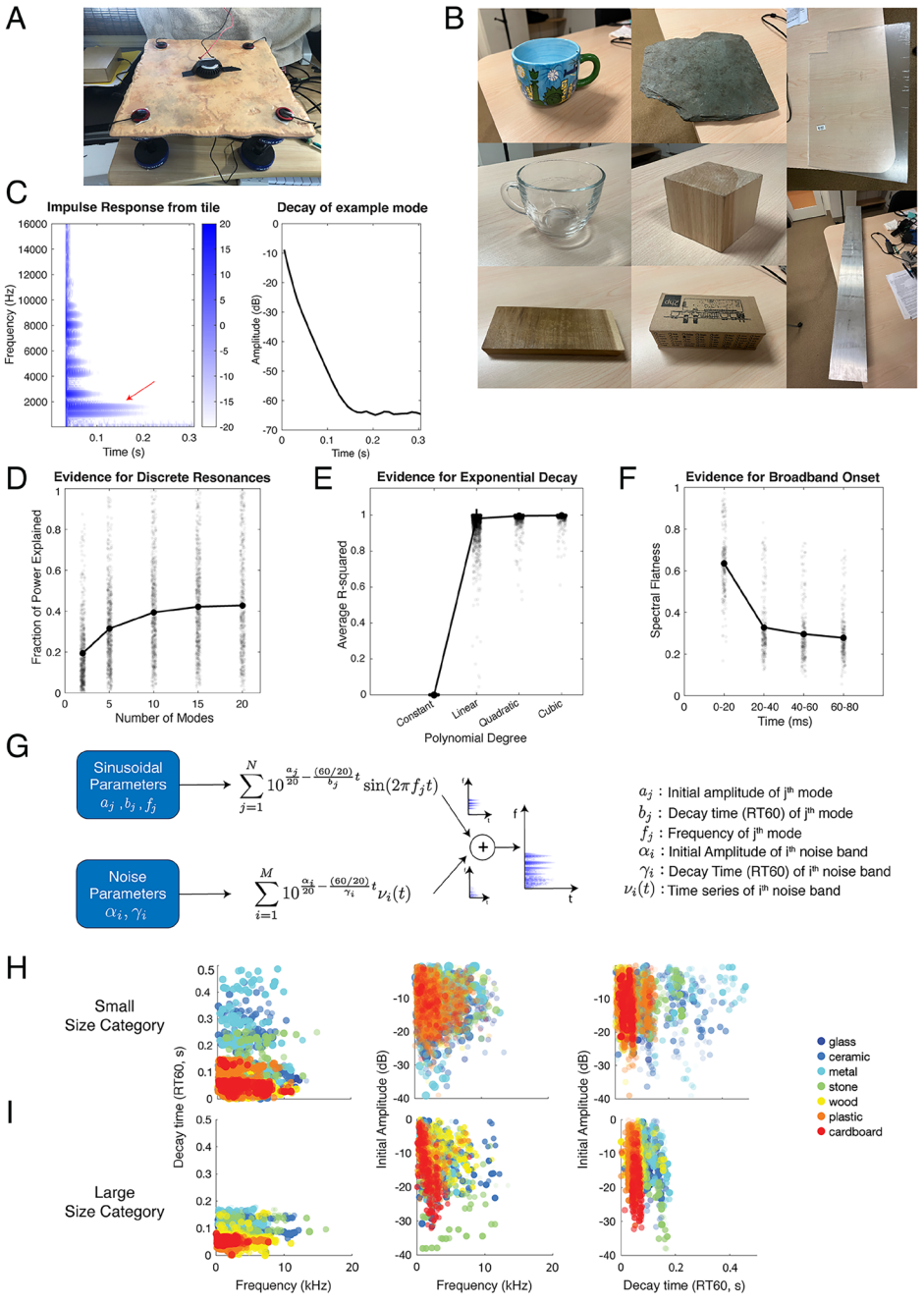



**Fig. 2.** Object impulse response measurement and statistics. (*A*) Apparatus for measuring object impulse responses. A noise signal is broadcast into an object with a contact speaker (in the center of the tile in the photo), and rerecorded from the object's surface via contact microphones (in the corners of the tile in the photo). The object's impulse response is estimated from the cross-correlation of the original noise signal and its rerecording. (*B*) Example objects from impulse response survey. (*C*) Spectrogram of example object impulse response (*Left*; color scale denotes amplitude in dB) and the energy decay of the mode indicated by the arrow (*Right*). (*D*) Evidence for discrete resonant frequencies. Power accounted for by the strongest modes of each impulse response. The contribution of sinusoidal modes to the impulse response can be accounted for by a fairly modest number of modes. Here and in *E* and *F*, dots plot results for individual object impulse responses (horizontal position of each dot is randomly jittered for ease of visibility). (*E*) Evidence of exponential decay of impulse response modes. Graph plots variance in the log-energy vs. time of each mode of each impulse response explained by polynomials of different orders. Analysis was restricted to the energy above the estimated noise floor. Most variance is accounted for by a linear function, indicative of exponential decay. (*F*) Evidence for broadband onset of impulse response. Graph plots spectral flatness of the power spectrum computed from successive 20-ms excerpts of impulse responses. High values of spectral flatness indicate a broadband spectrum. (*G*) Parametric model of object impulse response. Impulse response was modeled as a sum of decaying sinusoids (representing the object resonant modes) and decaying noise bands (used to model the onset of the impulse response, which was broadband). The model was fit to each recorded impulse response to obtain estimates of the mode parameters for each object. (*H*) Mode parameters for all objects in the Small size category (10 to 30 cm). Each dot plots a single mode from a single impulse response, color-coded by the object's material. (*I*) Mode parameters for all objects in the Large size category (1 to 3 m). Same conventions as *H*.

Downloaded from https://www.pnas.org by 86.130.54.74 on November 6, 2026 from IP address 86.130.54.74.

(for the object shown in Fig. 2*A*) illustrates the key properties. First, consistent with classical theory, nearly all impulse responses exhibited a set of relatively discrete frequencies—the resonant modes of the object—evident as horizontal stripes in the spectrogram. The mode frequencies are inharmonic, unlike some other natural sounds (26). Second, the power in these modes decays over time, closely approximated by exponential decay (Fig. 2 *C*, *Right* panel). Third, the impulse response typically begins with a broadband onset (i.e., a "click"), evident as the vertical stripe at the beginning of the spectrogram. This confirms the theoretical expectation that an impulse initially excites all frequencies, most of which decay very rapidly, leaving energy only in the resonant modes (5, 27).

We substantiated these observations quantitatively in the full set of impulse responses. First, we measured the fraction of impulse response power accounted for by the N strongest spectral components (Fig. 2*D*). Most of the power contributed by sinusoidal modes was due to the 10 strongest modes. Thus, most impulse responses are dominated by a small set of relatively discrete modes, though the strength of these resonant modes (relative to the rest of the impulse response power, which we will model with noise) varied considerably across objects. Second, we fit polynomials to the log-energy decay in these 10 modes. Most of the variance in the decay (excluding the noise floor) was accounted for by a linear function (i.e., an exponential function on a log scale), demonstrating that the modes nearly always decay exponentially (Fig. 2*E*). This result allows each mode to be summarized with three parameters: the frequency, the decay time, and the starting amplitude. Third, we measured the spectral flatness of the impulse responses in short time windows (Fig. 2*F*). The initial time window consistently had a more broadband spectrum than later windows, indicating a broadband onset to the impulse response. These findings are consistent with classical theory (2, 3) and smaller scale measurements of object sounds (5), but show that the theory holds broadly for everyday objects, and that the acoustic properties of everyday object have a fairly stereotyped form as a result.

**Parametric Model of Object Impulse Responses.** The stereotyped form of object impulse responses can be captured with a parametric model (Fig. 2*G*) that sums sinusoidal resonant modes and decaying noise bands (critical for correctly capturing the onset). We fit the parameters of this model to each measured object impulse response, using gradient-based optimization (*Materials and Methods*). This yielded a description of each impulse response in terms of its mode frequencies, decay rates, starting amplitudes, and noise parameters.

**Different Impulse Responses From an Object Are Similar.** For each object we measured multiple impulse responses with different surface microphone locations. Each of these impulse responses summarizes the sound that would radiate from a point on the object's surface given a force applied at the speaker location. Impulse responses from the same object tended to share a considerable number of modes (*SI Appendix*, Fig. S1; see *Materials and Methods* for details). This finding suggests that a single impulse response is likely to be representative of the modal properties of an object, and thus is also likely to be representative of the impulse response for an external point of measurement that would combine the sound from different points on an object's surface.

**Object Impulse Response Statistics—Effects of Material and Size.** We observed pronounced effects of material, and to a lesser extent, object size. Fig. 2 *G* and *H* illustrate these effects in the Small (10 to 30 cm) and Large (1 to 3 m) size categories (see *SI Appendix*, Figs. S2–S4 for all size categories). Mode properties varied with material [harder materials like metal and ceramic show longer decay rates and higher frequency modes, consistent with prior theory (10–12, 14, 28)]. There is also some variation with size (larger objects tend to have lower frequency modes, consistent with prior theory, and shorter decay times, plausibly due more external damping of larger objects than smaller objects; *SI Appendix*, Fig. S5). The effects of material were fairly consistent across object sizes (Fig. 2 *G* and *H* and *SI Appendix*, Fig. S2), and the effects of size were fairly consistent across material (*SI Appendix*, Figs. S3–S5). The distributions for different materials and sizes nonetheless overlap substantially.

**Generative Model of Impulse Responses.** We built a generative model of impulse responses (Fig. 3*A*) by fitting multivariate Gaussian distributions to the three parameters of a mode (frequency, starting amplitude, and decay time) and the two parameters of the noise bands (starting amplitude, decay time) conditioned on material and size. The fitted distributions reflect the effects of

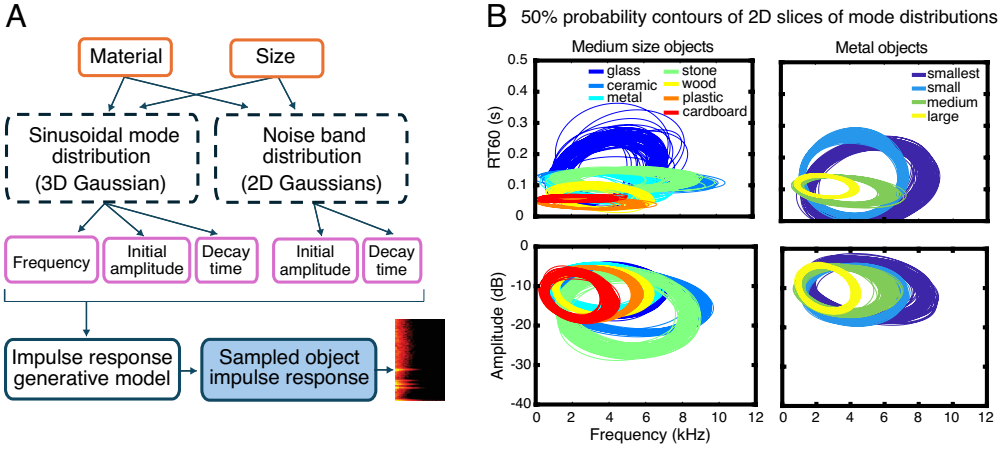



**Fig. 3.** Generative model of impulse responses. (*A*) Mode parameter dependencies were modeled with multivariate Gaussian distributions conditioned on object material and size. (*B*) Constant probability ellipsoids (half-maximum of density) from fitted Gaussian distributions. To illustrate the stability of the distribution estimates, we plot ellipsoids from Gaussians fit to different bootstrap samples of the impulse responses for each material and size category. Gaussian distributions were truncated at zero for all dimensions. Harder materials (glass, ceramic, metal, and stone) tend to have modes with higher frequencies and longer decay times. Larger objects tend to have lower frequency modes. Object size ranges were <10 cm (Smallest), 10 to 30 cm (Small), 30 to 100 cm (Medium), and 1 to 3 m (Large).

Downloaded from https://www.pnas.org by 86.130.54.74 on November 6, 2026 from IP address 86.130.54.74.

material and size evident in the scatter plots of Fig. 2 *G* and *H*, but also reveal covariance between the three mode parameters, evident in constant-probability ellipsoids that are oriented in some cases (Fig. 3*B*). See *SI Appendix*, Fig. S6 for the distribution parameters for the noise bands. Unlike some methods of generating object resonances (6–8), the role of physical variables here is statistical, specifying signal distributions rather than simulating vibrations. The generative model can be used to sample impulse responses (by sampling parameters for sinusoidal modes and noise bands). It can also be used to measure the likelihood of an impulse response under the real-world distribution, which we will use later in the paper.

**Physics and Acoustics of Impact Force.** The other essential ingredient in impact sound generation is the force created as objects come into contact, compress against each other, and then push apart. We computed contact force from an extension of a classical spring model (29). This formulation yields a contact force as a half-cycle of a sinusoid, with the spring constant determined by the mass of the impacting object and the stiffnesses of both objects (Fig. 4*A*). Because the typical duration of the force is less than a millisecond, the primary acoustic effect is to filter the object impulse response, with the transfer function shown in Fig. 4*B*. The overall amplitude of this transfer function (and thus the resulting sound) increases with the ball mass, as expected. Moreover, the peak frequency decreases for heavy balls and/or softer materials, as both of these increase the time the surface and ball are in contact, i.e., the width of the half-cycle sinusoid, resulting in a more low-pass filter. The formulation thus encompasses the two main acoustic cues previously associated with mass (amplitude and spectral centroid) (16, 17). Recorded impact sounds generated by dropping pellets of different masses onto object surfaces showed the expected effects of object mass (Fig. 4*C*), with heavier objects preferentially exciting lower frequency modes (Fig. 4*D*). To generate the force for an impact sound, we set the object stiffnesses to material- and size-specific values (*Materials and Methods*).

**Generative Model of Impact Sounds.** By combining the models of impulse responses and contact forces, we obtain a generative model of impact sounds (Fig. 5*A*). The model generates a sound from a small set of physical variables (the object mass, material, size, and speed of impact), providing a tool we can use to synthesize stimuli and study perception.

**Experiment 0: Validation of Impulse Response Recordings and Impact Sound Synthesis.** We validated the resulting synthetic sounds with an experiment measuring material identification using recorded and synthetic sounds. Participants heard a single impact sound and categorized the material into one of seven categories (Fig. 5*B*). We compared material recognition for recorded impact sounds produced by a small pellet dropped on an object, sounds rendered from the generative model using recorded impulse responses, and sounds rendered from the generative model using sampled impulse responses. We used objects from the middle two size categories (10 to 30 cm and 30 to 100 cm), as these were most amenable to recording clean impacts (*Materials and Methods*).

We ran experiments online to facilitate large sample sizes. Prior research generally indicates that online experiments replicate in-lab experiments (30–32) provided measures are taken to ensure

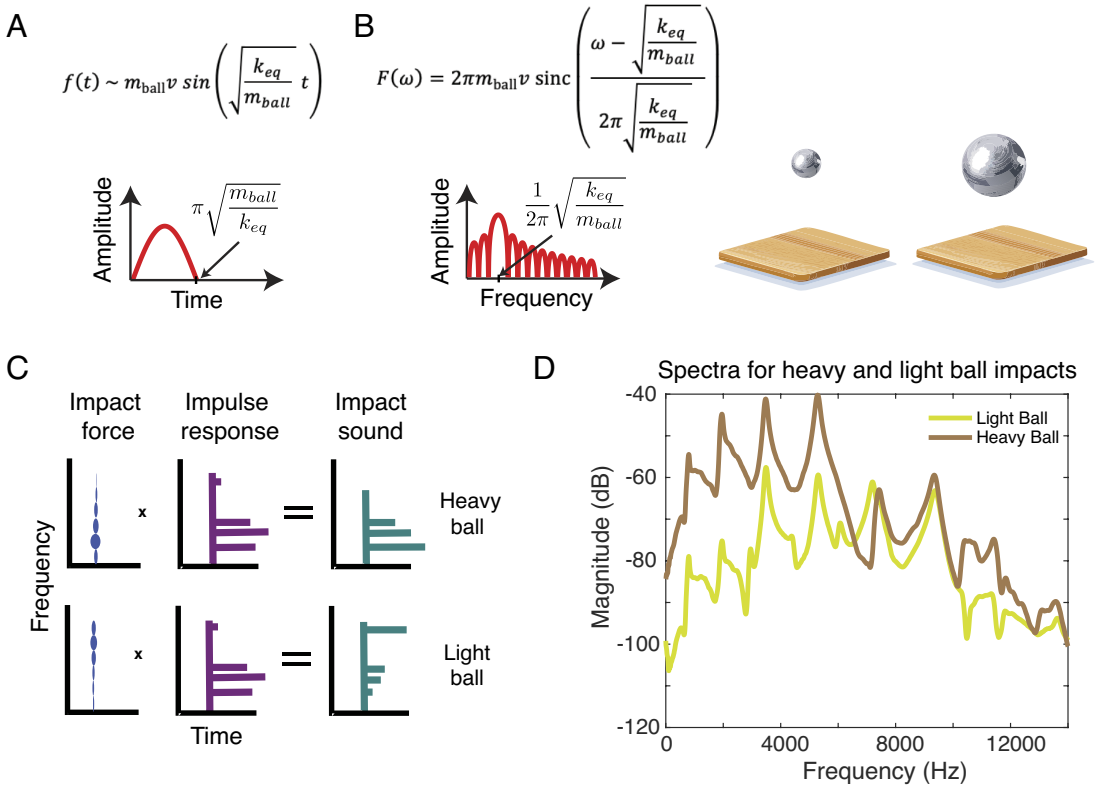



**Fig. 4.** Effects of impact force. (*A*) Acoustic effects of impact force. The force produced by an impact increases and then decreases over time as the objects come into contact, compress, and then move apart. The forcing function is approximately a half-cycle of a sinusoid, where $\omega$ is frequency, $k_{eq} = \frac{k_{ball} k_{surface}}{k_{ball} + k_{surface}}$ is the equivalent spring constant (i.e., resistance to compression) of the combined ball and surface system, $v$ is the velocity of the ball at impact, and $m_{\text{ball}}$ is the mass of the ball. (*B*) The effect of the forcing function in the frequency domain is given by its transfer function. (*C*) Schematic depiction of the effect of impacting object mass on the resulting impact sound. The peak of the transfer function varies with mass, filtering the impacted object impulse response to produce an impact sound. (*D*) Power spectra of recorded impact sounds from different impacting object masses. See *SI Appendix*, Fig. S7 for further validation.

Downloaded from https://www.pnas.org by 86.130.54.74 on November 6, 2026 from IP address 86.130.54.74.

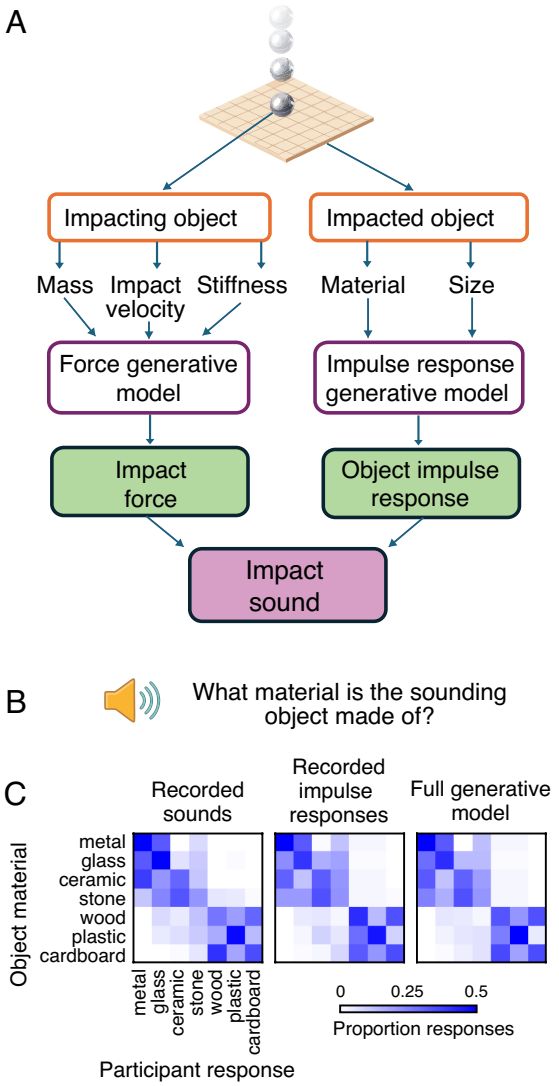



**Fig. 5.** Generative model of impact sounds and its validation. (*A*) Generative model of impact sounds. The full generative model combined the generative models for impact sounds and contact force shown in Figs. 3 and 4. (*B*) Experimental task for Experiment 0. Listeners heard a sound and selected the object material from a list of seven choices. (*C*) Confusion matrices for sounds from each experimental condition. We note that this experiment included only two of the four size categories and used material-specific stiffness values in the synthetic sounds, and it is thus unsurprising that the pattern of confusions is somewhat different from that measured in Experiment 2.

standardized sound presentation (33) and participant compliance. See *SI Appendix*, Figs. S8–S11 for evidence that results are similar for in-ear vs. over-ear sound presentation.

Humans were well above chance at identifying an object's material from sound, and performed similarly across the three conditions (recorded sounds: 38.5% correct; sound synthesized from recorded impulse responses: 37.7% correct; sounds synthesized from the full generative model: 38.2% correct). Humans nonetheless made many errors (Fig. 5*C*), consistent with prior studies (12, 13, 15). These errors plausibly reflect overlap between distributions of impulse responses of different materials, evident in Figs. 2 and 3 and *SI Appendix*, Figs. S2–S4. The confusion matrices were strongly correlated (recorded sounds vs. sounds synthesized from recorded impulse responses: $r = 0.94$, $P < 0.001$; recorded sounds vs. sounds synthesized from the full generative model: $r = 0.95$, $P < 0.001$). These results provide evidence that the assumptions underlying the generative model and impulse response recordings are sufficiently accurate as to produce synthetic sounds that are perceived in much the same way as real-world sounds.

**Lesioned Variants of Generative Model.** We investigated the perceptual importance of the physical regularities built into the model using "lesioned" model variants that each violated some aspects of the full generative model (Fig. 6*A*). These variants were chosen to deviate from the real-world impulse response distribution to varying extents. We quantified this deviation by measuring the log-likelihood under the real-world distribution of impulse responses sampled from each condition (Fig. 6*B*). All model variants produced impulse responses that were less likely under the real-world distribution (compared to Recorded and Full model impulse responses), but the magnitude of this effect varied across conditions, as intended.

**Experiment 1: Realism of Synthesized Impact Sounds.** We first tested whether the realism of sounds produced by our generative model depended on the physical and statistical regularities captured by the model. Participants heard two sounds and judged which was from a real object (Fig. 7*A*). One of the sounds was generated from a recorded impulse response, and the other was generated using the generative model or one of the model variants in Fig. 6*A*.

Human listeners were unable to distinguish sounds from the generative model from those made from recorded impulse responses (Fig. 7*B*; $t(86) = 0.420$, $P = 0.68$; $d = 0.046$). However, sounds from the model variants (each of which violated a physical or statistical regularity found in real-world objects) were distinguishable from those from the Recorded condition. The effect size depended on the variant, but even the most subtly different condition (Inverted covariance) was significantly distinguishable [$t(86) = 4.82$, $P < 0.001$, statistically significant after Bonferroni correction; $d = 0.50$; see *SI Appendix*, Table S5 for exhaustive statistical reporting]. This result suggests that our model successfully captures the perceptually important features of real-world impacts and that this success depends on its physically inspired structure and empirically captured statistical regularities.

**Experiment 2: Material Recognition of Synthesized Impact Sounds.** As a second test of the generative model, we tested the effect of model lesions on material perception. Participants heard a single impact sound and categorized the material into one of seven categories (same task as Experiment 0; Fig. 7*C*). The sound was generated using a recorded impulse response, using the generative model, or using one of the model lesion variants. Here and in subsequent experiments, we used a single stiffness value (unlike in Experiments 0 and 1), so as to isolate the effect of the object impulse response.

Our analysis compared material confusion matrices across conditions. The confusions produced by actual object impulse responses were closely mirrored by those for samples from the full generative model (Fig. 7*D*). We note that the confusions differed somewhat from those of Experiment 0 (Fig. 5*C*); these differences are unsurprising given differences in the stimuli (Experiment 2 eliminated material cues from the forcing function and sampled from all four size categories). More notably, confusions for each of the model variants deviated from the pattern seen for the full generative model. The correlation between the matrix for recorded impulse responses and that for model-generated sounds was close to 1 for sounds from the full generative model but was lower for each of the model variants ($P < 0.003$ for each paired comparison, via permutation test, significant after Bonferroni correction; difference in correlation ≥0.12). This result indicates that the generative model captures the acoustic regularities underlying human

Downloaded from https://www.pnas.org by 86.130.54.74 on November 6, 2026 from IP address 86.130.54.74.

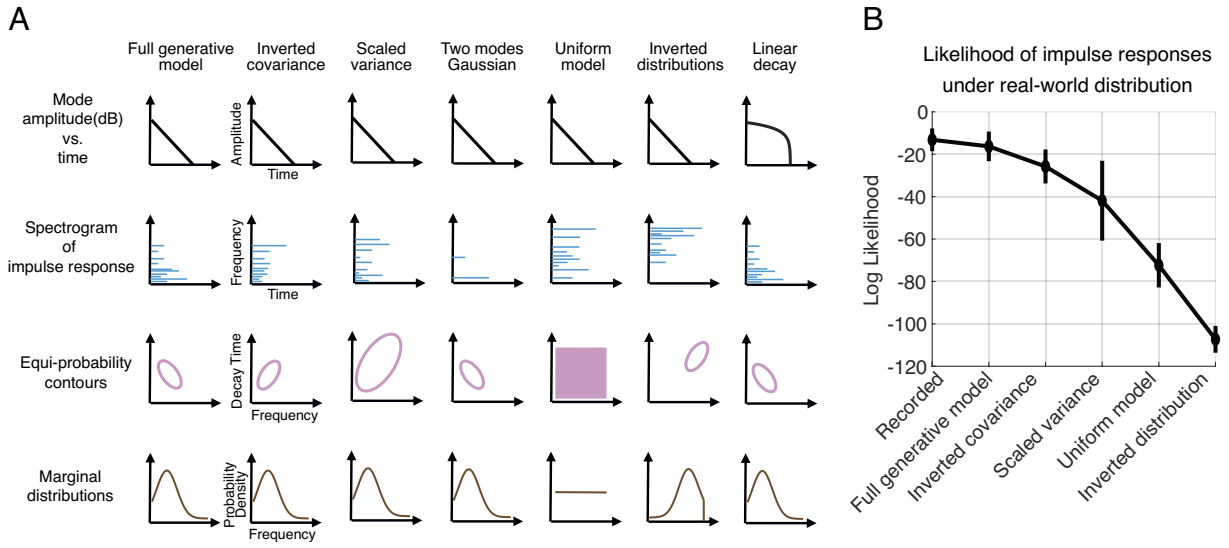



**Fig. 6.** Lesioned variants of generative model. (*A*) Variants of the generative model used in perceptual experiments. The top row shows the energy decay of an example sinusoidal mode, which was exponential in the full generative model (appearing linear on a dB scale) but that was altered to be linear in one of the model variants (appearing concave down on a dB scale). The second row shows schematic spectrograms of impulse responses. The full generative model produced impulse responses that have 10 modes, but this was reduced to 2 modes in one of the model variants. The third row shows equiprobability contours for the distributions from which mode parameters were sampled. In the full generative model these distributions were Gaussian, with covariances measured from real-world impulse responses. In two of the model variants the covariance was inverted, producing the opposite relationship between mode parameters (shown here for frequency and decay time, also evident in the schematic spectrogram); in one of these variants the variance was also scaled up. In two other model variants the distributions took a different form: uniform or "inverted" (having high probability where the full model has low probability, and vice versa). The fourth row shows distributions marginalized over decay time and amplitude. Note that Gaussian distributions were truncated at zero on the frequency axis. (*B*) Log-likelihood of impulse responses from model variants, evaluated under the full generative model. We omitted the Two modes and Linear decay conditions as their mode parameters were taken from the full model.

material recognition from sound and that the lesions interfere with human material perception.

Overall, Experiments 1 and 2 confirm that humans are sensitive to the acoustic structure of impact sounds and associate this structure with object properties. The experiments also establish that our generative model captures much of the relevant acoustic structure.

**Mass Discrimination Within and Across Impacted Surfaces.** One of the most interesting, and most neglected, aspects of the perception of physical interactions is the question of whether and how humans can solve the ill-posed problems created by the confluence of physical variables from multiple objects. We investigated this question in two scenarios (Experiments 3 and 4, respectively). The first scenario was that of estimating the mass of a dropped object. This problem is in general ill-posed because the mass of the dropped object and the impulse response of the impacted surface both influence the sound that is produced (Figs. 1 and 4). To study this problem, we designed the task shown in Fig. 8 *A*–*C*, in which a heavy ball and a light ball are dropped onto surfaces. The task was to determine which ball is heavier. We asked whether performance depended on whether the impulse response of the impacted surface was drawn from the natural distribution (which limits inference of the impacted surface material, as shown in Experiment 2).

If the two balls are dropped on the same surface from the same height (Fig. 8*A*), mass discrimination can be straightforwardly performed by comparing the amplitude and frequency content of the resulting sound (the heavier ball will produce a higher-amplitude and lower- frequency sound; Fig. 4 *B*–*D*). However, if the two surfaces are of different materials (Fig. 8*B*), the task becomes ill-posed. Because the power and frequency content of impact sounds are jointly determined by the impact force and object impulse response (Fig. 8*D*), a heavier object dropped on a harder surface (e.g., ceramic) and a lighter object dropped on a softer surface (e.g., plastic) might be difficult to distinguish simply via amplitude and frequency content.

We hypothesized that listeners might leverage implicit knowledge of impulse response regularities to overcome the ill-posedness. For instance, knowing that harder objects tend to have modes that are both higher frequency and that have slower decay rates (Fig. 2 *H* and *I*) might enable a listener to discount the contribution of the impulse response to the sound spectrum, enabling estimation of the mass. This hypothesis predicts that mass inference should be impaired if the impulse response of the struck surface deviates from the naturally occurring impulse response distribution, since a listener's implicit knowledge of impulse responses would then be less useful in factoring out the impulse response from the force.

**Experiment 3a: Mass Discrimination Within the Same Impacted Surface.** We first tested human listeners in an experiment in which the same impulse response was used for the two sounds, simulating the setting in which the two balls are dropped on the same surface (Experiment 3a; Fig. 8*A*). In this case, the effects of the impulse response do not have to be separated from that of the impact force; mass can be discriminated simply by comparing the acoustic cues that are correlated with mass (e.g., amplitude and spectral centroid; Fig. 4). Moreover, the trial structure was intended to make these cues obvious to the listener (Fig. 8*E*): the two sounds on a trial contain the same frequencies (object resonances), such that there is a single factor of variation distinguishing the pairs of stimuli. If the different stimulus conditions (synthesized from different generative model variants) preserve these individual cues as intended, performance should be comparable for all conditions.

As shown in Fig. 8*G*, performance was indeed similar for all model variants (statistics provided below). This result indicates that when the problem is not ill-posed, listeners can discriminate mass using cues that are preserved by all the model variants.

Downloaded from https://www.pnas.org by 86.130.54.74 on November 6, 2026 from IP address 86.130.54.74.

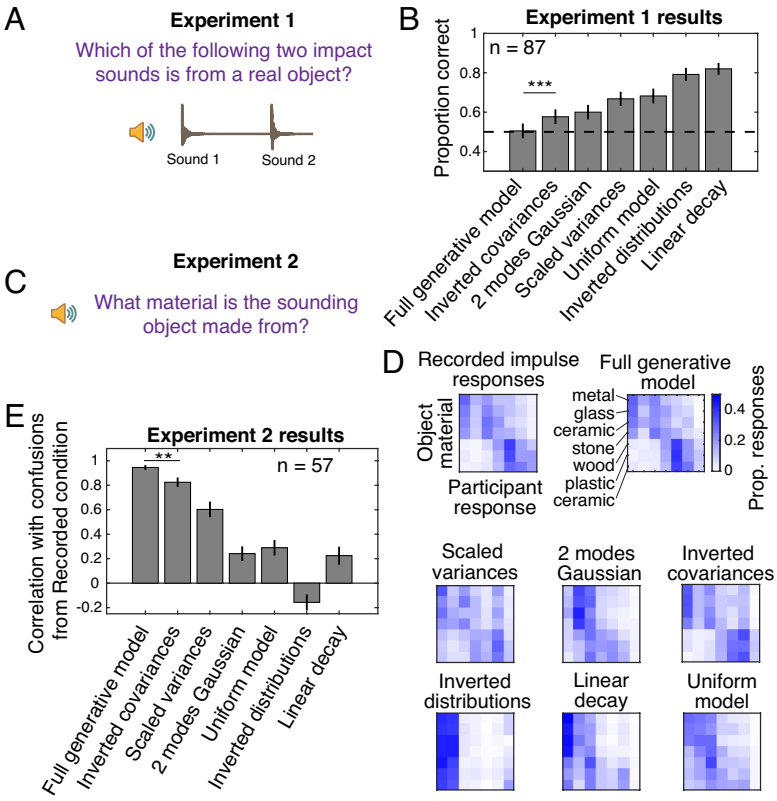



**Fig. 7.** Realism of synthetic impact sounds depends on physical and statistical regularities. (*A*) Experiment 1 task. Listeners heard two sounds and judged which was from a real object. One sound was generated from a recorded impulse response, and the other was generated from a synthetic impulse response sampled either from the generative model or one of several "lesioned" model variants. (*B*) Results of Experiment 1. Participants were unable to distinguish sounds synthesized from the generative model from those measured from actual objects, but each of the lesioned models was distinguishable to some extent. Error bars plot SEM, obtained via bootstrap. Dashed line plots chance performance level. *** denotes $P<0.001$. (*C*) Experiment 2 task. Listeners heard a sound (same conditions as Experiment 1) and selected the object material from a list of seven choices. The sound was either generated from a recorded impulse response or from a synthetic impulse response sampled from the generative model or a lesioned version of the model. (*D*) Confusion matrices for sounds from each experimental condition. We note that this experiment used different stimuli from Experiment 0 (it included all four size categories, and used a single stiffness value for all materials, for ease of comparison with Experiments 3 and 4), and thus unsurprisingly produced a different pattern of confusions. (*E*) Correlation between confusion matrix for sounds from recorded impulse responses and the matrix for each model condition. Error bars plot SEM, obtained via bootstrap. ** denotes $P<0.01$.

**Experiment 3b: Mass Discrimination Across Different Surfaces.** We then tested human listeners in an experiment in which the two impact sounds were synthesized from different impulse responses, as if the balls were dropped on different surfaces, rendering the discrimination ill-posed (Experiment 3b; Fig. 8 *B, D,* and *F*). Unlike Experiment 3a, there is not a single acoustic dimension that clearly distinguishes the two stimuli on a trial. In this case, we observed large differences between conditions (Fig. 8*H*). Performance improved as the simulated mass difference between the two balls increased but was worse for the five conditions that violated real-world impulse response regularities.

We quantified performance as the threshold mass ratio required to yield 80% correct (Fig. 8 *I* and *J*). These thresholds were similar across conditions for Experiment 3a (same object; only the Inverted distribution was significantly worse than the Recorded condition, $P = 0.001$; threshold ratios for other conditions ranged from 1.00 to 1.13, $P > 0.1$), but varied across conditions for Experiment 3b (different objects). In particular, the two conditions with realistic impulse responses yielded similar thresholds ($P = 0.27$, via permutation test; ratio between thresholds = 1.04), but each of the five conditions from the alternative generative models yielded significantly worse performance ($P \leq 0.001$ for each paired comparison with the Full model condition, statistically significant after Bonferroni correction; ratio between thresholds ≥ 2.15). Moreover, in the conditions of Experiment 3b in which the impulse response statistics were varied, the thresholds covaried with the average log-likelihood of the impulse responses (Fig. 8 *K* and *L*), consistent with the idea that humans make use of a prior over impulse responses to separate out the impulse response from the impact force that conveys mass. This result confirms the key prediction of the hypothesis that listeners overcome the ill-posed nature of the problem by leveraging implicit knowledge of the statistical and physical regularities of sound generation.

**Results not Explained by Individual Acoustic Cues.** To further confirm that the differences in performance across conditions could not be explained by individual acoustic cues, we implemented two "cue discriminator" models that performed the task by comparing either the spectral centroid or amplitude of the two sounds, and picking the lower or higher amplitude sound, respectively. These two cue discriminator models performed similarly across conditions (Fig. 9 *A–C*), as expected. Moreover, the centroid discriminator underperformed humans for the two naturalistic conditions (Recorded and Full model; Fig. 9*C*). These results indicate that all stimulus conditions contain similar low-level acoustic cues, as intended, and underscore that humans are not performing the task with a simple comparison of these cues.

**Experiment 3c: Mass Discrimination Across Surfaces with Level-Equated Sounds.** As an additional test of the role of the acoustic cues associated with mass (amplitude and spectral centroid), we ran a control experiment (Experiment 3c) in which the amplitude of the two sounds was equated. In this case, performance was worse overall but was still well above chance for the two conditions that were otherwise realistic (Recorded and Full model), while notably worse for the conditions from lesioned model variants (Fig. 9*D*; significantly worse than Recorded in all cases, $P < 0.001$, permutation test). This task obviously cannot be performed by the amplitude cue discriminator (Fig. 9*E*) and is also not explained by the centroid discriminator (Fig. 9*F*), which is more accurate than humans for large mass-ratios (>10) but less accurate for low mass-ratios (<10), and which again does not differ between conditions (unlike humans). We note that the relative sound levels here were physically incorrect, which in some contexts impairs human recognition (32), and could explain the worse performance relative to the realistic case where there are level differences between sounds. This result nonetheless reinforces the idea that listeners are not basing judgments on amplitude cues alone.

Downloaded from https://www.pnas.org by 86.130.54.74 on November 6, 2026 from IP address 86.130.54.74.

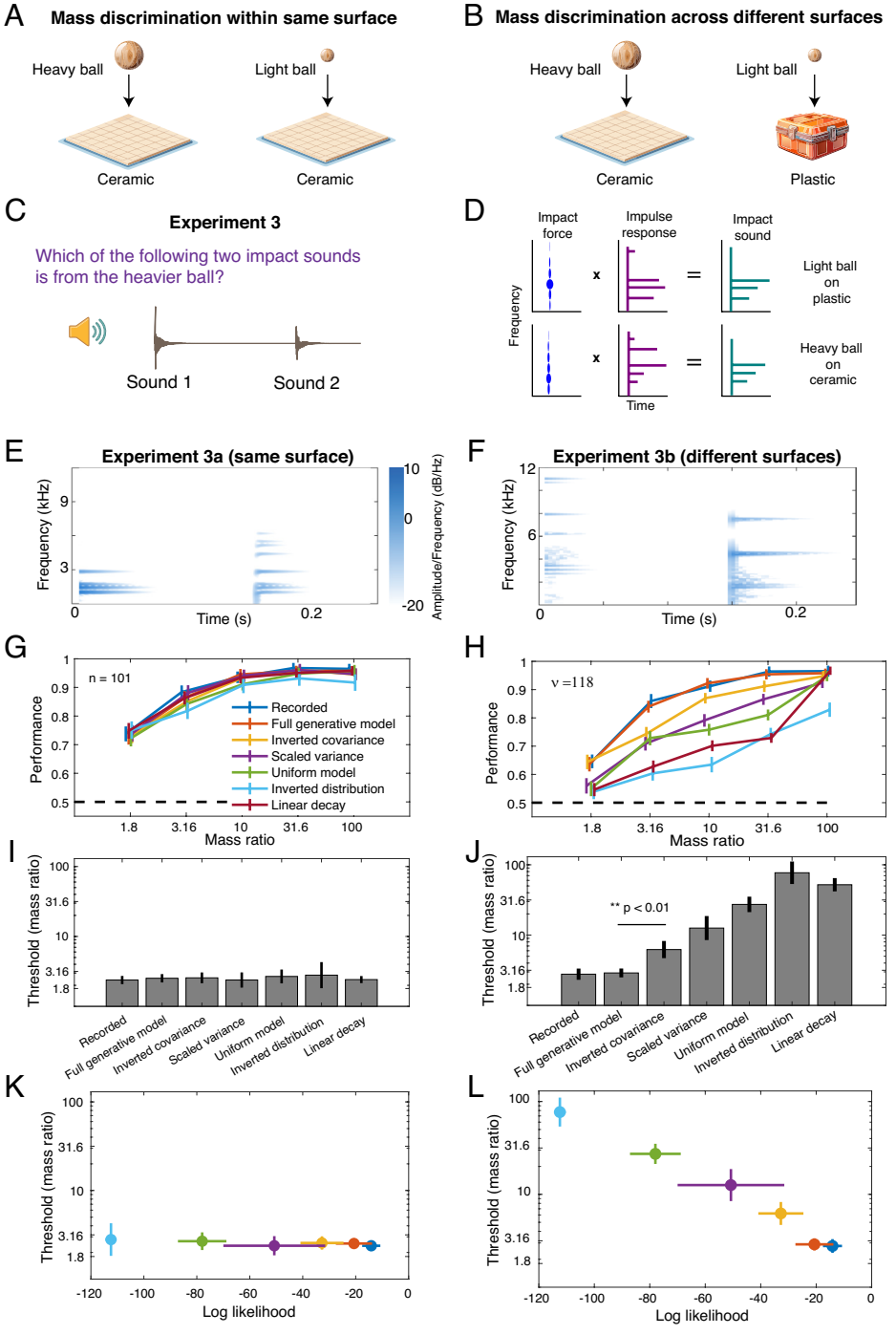



**Fig. 8.** Mass discrimination within and across impacted surfaces (Experiment 3 a and b). (*A* and *B*) Experiment scenario: Two balls of different masses are dropped on either the same object, or on two different objects. (*C*) Experiment 3 task. Listeners heard two impact sounds and judged which was produced by the heavier ball. The experiment manipulated the difference between the two masses and the extent to which the impact sounds were faithful to the statistical regularities of real-world object impulse responses. (*D*) The joint influence of ball mass and surface material on the spectrum of the impact sound renders the inference of mass ill-posed. (*E*) Spectrogram of stimuli on example trial of Experiment 3a, in which both sounds are rendered from the same metal surface. The object resonances (modes) are the same in the two stimuli (note that the three prominent frequencies in the first stimulus are also present in the second). As a result, the discrimination can be performed using simple acoustic cues of level and centroid (e.g., in the first stimulus, the peak amplitude is higher and the spectral centroid is lower). (*F*) Spectrogram of stimuli on example trial of Experiment 3b, in which one object is ceramic and the other is stone. Here, the object resonances (modes) are different in the two stimuli (note that each stimulus contains a different set of frequencies). As a result, the simple acoustic cues that are diagnostic in Experiment 3a are not as diagnostic for solving the task. (*G*) Results of Experiment 3a: proportion correct plotted as a function of the mass ratio between the two balls. Here and in *H*–*J*, error bars plot SEM, obtained via bootstrap. Dashed line plots chance performance level. (*H*) Results of Experiment 3b. (*I* and *J*) Discrimination thresholds for Experiments 3a and 3b (the mass ratio resulting in 80% correct, estimated by fitting splines to the data in *D*). (*K* and *L*) Comparison of discrimination thresholds from Experiments 3a and 3b with log-likelihood of impulse responses under full generative model. The linear decay condition is omitted as the impulse response parameters were taken from the Full model, and thus do not reflect the deviant features of that model variant. Error bars on thresholds plot SEM, derived via bootstrap. Error bars on log likelihood plot SD across impulse responses.

Overall, the results indicate that listeners are able to separate the contribution of object resonances from that of the force exerted between objects, in this case as is determined by the mass of the impacting object, and that they rely on knowledge of real-world impulse response regularities to do so.

**Experiment 4: Perception of Damping.** A second everyday scenario in which perceptual inferences are ill-posed involves the perception of external damping (28). External damping occurs when one object abuts another, absorbing some fraction of the other object's vibrations. External damping is distinct from the internal damping in which energy dissipates within an object at rates that depend on the material [reflected in the material-dependent decay rates seen in our impulse response measurements, and in prior theoretical accounts of material perception (10, 12, 14, 28)]. In this paper, we use the term damping to refer to external damping.

Nearly all objects are externally damped to some extent, as gravity usually causes them to rest on something, but some conditions produce much more damping than others. Consider, for instance, the sound of tapping a coffee mug sitting on a table vs. held firmly in someone's hands. Anecdotal experience suggests we are able to discern the degree of external damping to some extent, but the acoustic basis of any such abilities was unclear. We measured the acoustic effects of external damping on object resonances by clamping different fractions of a board and then measuring its impulse response (Fig. 10 *A* and *B*). The primary effect of damping was to lower the decay time of the object resonances captured by

Downloaded from https://www.pnas.org by 86.130.54.74 on November 6, 2026 from IP address 86.130.54.74.

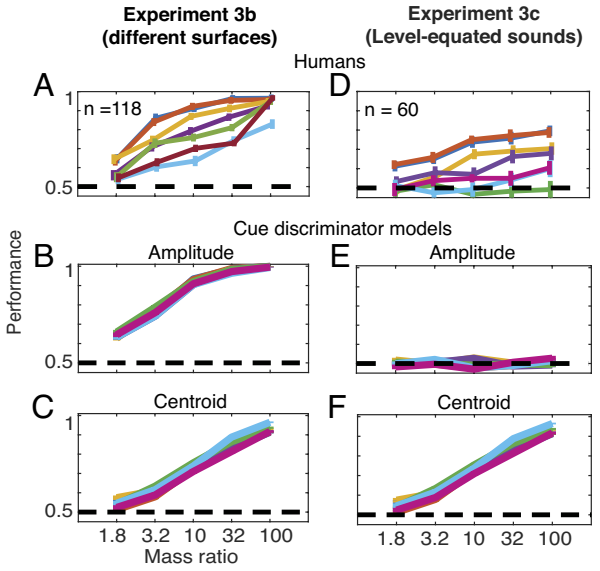



**Fig. 9.** Human mass discrimination is not explained by single acoustic cues. (*A*) Results of Experiment 3b, replotted for comparison with other panels. Here and elsewhere, error bars plot SEM, obtained via bootstrap. (*B*) Performance of amplitude cue discriminator model on Experiment 3b. (*C*) Performance of spectral centroid discriminator model on Experiment 3b. (*D*) Results of Experiment 3c (identical to Experiment 3b except that the peak amplitude of each sound was set to a fixed reference, such that the two sounds on each trial were level-equated). The worse performance compared to Experiment 3b is unsurprising as the resulting stimuli are physically incorrect. (*E*) Performance of amplitude cue discriminator model on Experiment 3c. (*F*) Performance of spectral centroid discriminator model on Experiment 3c.

the object's impulse response (see *SI Appendix*, Fig. S12 for additional measurements).

This acoustic effect raises the possibility that it could be challenging to disentangle the effects of external damping from the effects of material, because both of them influence object resonance decay times, potentially making the inference of damping ill-posed (Fig. 10*C*). However, with prior knowledge of the distribution of impulse responses, one might be able to use the frequencies of the resonances to estimate the expected decay rates in the absence of damping, enabling the effect of material to be factored out and thus damping to be more accurately estimated.

To test this idea, we had participants listen to single impact sounds and judge whether the object making the sound was resting freely on a table or held tightly in someone's hands (Fig. 10*D*). We simulated different degrees of damping by decreasing the mode decay times by different proportions. Because this was a single-interval task, performance was quantified as d-prime.

As shown in Fig. 10*E*, participants were able to determine whether an object was externally damped, but were substantially better for the two conditions with naturalistic impulse responses. When summarized with a threshold for each condition (here the amount of simulated damping needed to produce a d-prime of 1 in each condition; Fig. 10*F*), the four unnatural conditions all produced significantly worse performance than the natural conditions ($P \leq 0.003$ for each paired comparison with the Recorded condition, via permutation test, statistically significant after Bonferroni correction; threshold ratio ≥1.23). As in Experiment 3, thresholds varied inversely with the likelihood of the impulse responses under the real-world distribution (Fig. 10*G*). These results indicate that listeners rely on implicit knowledge of the distribution of object resonance properties to be able to discern when an object is damped. This experiment thus provides another example of human abilities to separate distinct physical causes of sound, using internalized knowledge of the world.

## Discussion

We investigated human perceptual inferences of physical variables from sound, leveraging measurements of real-world object acoustics. We conducted the first large-scale set of object impulse response measurements, revealing systematic regularities in the way that objects of different materials and sizes produce sound when struck. We then developed a physically constrained generative model using these measurements. The model allowed us to sample and synthesize new instances of impact sounds that sounded realistic and that conveyed physical variables such as material and mass to human listeners, providing a tool with which to probe their perception. We validated the model with measurements of actual impact sounds and with perceptual experiments showing that the model components were sufficient and necessary to render realistic impact sounds. We then asked whether listeners show signs of separating the contributions of the distinct physical variables that contribute to any single sound. We found that human listeners could discern mass and damping despite variation in material, consistent with such a separation, but that they depended on implicit knowledge of physical regularities. In particular, when object impulse responses were drawn from alternative distributions, human performance degraded in a way that was predicted by the likelihood of the impulse responses under the estimated real-world distribution, indicating that listeners use knowledge of this distribution to factor out the effects of material when judging mass and damping. Notably, the ability to infer the properties of one of the objects in an event (the mass of the impacting ball, or the presence of a hand holding the struck object) was dependent on whether the properties of the other object could be reliably inferred from resonant modes.

The results suggest that the brain has internalized aspects of the physical and statistical structure by which sound is generated by objects, and performs inference that leverages this structure to overcome the intrinsic ill-posedness of the problem. The results implicate knowledge of the distribution of object resonances, but leave open the details of how this knowledge is applied. In principle, listeners might have a fairly rich model of the physical generative process in their heads, which they invert via Bayesian inference to estimate physical variables. However, our results are equally consistent with a wide range of alternative (approximate) inference schemes, such as discriminative or associative models that implicitly encode the distribution of object resonances without explicitly representing underlying physical causes. The results appear to rule out the commonly proposed idea that physical variables are identified with particular acoustic cues, as such accounts do not predict the strong dependence on whether object resonances conform to the natural distribution.

**Relation to Prior Work.** Our research builds on previous studies of human perception of impact sounds. Much of this work has focused on deriving individual cues from physical considerations and testing whether these cues can be used to make perceptual judgments (9–13, 15, 17–19). We built on this work in three respects. First, we addressed the ill-posed nature of the inference problem faced by observers in realistic settings in which multiple physical variables are unknown. This ill-posedness has been noted by others (28) but has only been studied in a few settings (12, 14, 16, 17), typically in the context of a search for cues to one variable that might be invariant to other variables. Our results suggest that human judgments are not exclusively driven by individual cues (at least not those that would naturally be associated with the variables of mass and damping that we studied), and instead involve a more sophisticated inference that to some extent factors out the different variables that contribute to a sound. Second, we developed a method to synthesize arbitrarily many

Downloaded from https://www.pnas.org by 86.130.54.74 on November 6, 2026 from IP address 86.130.54.74.

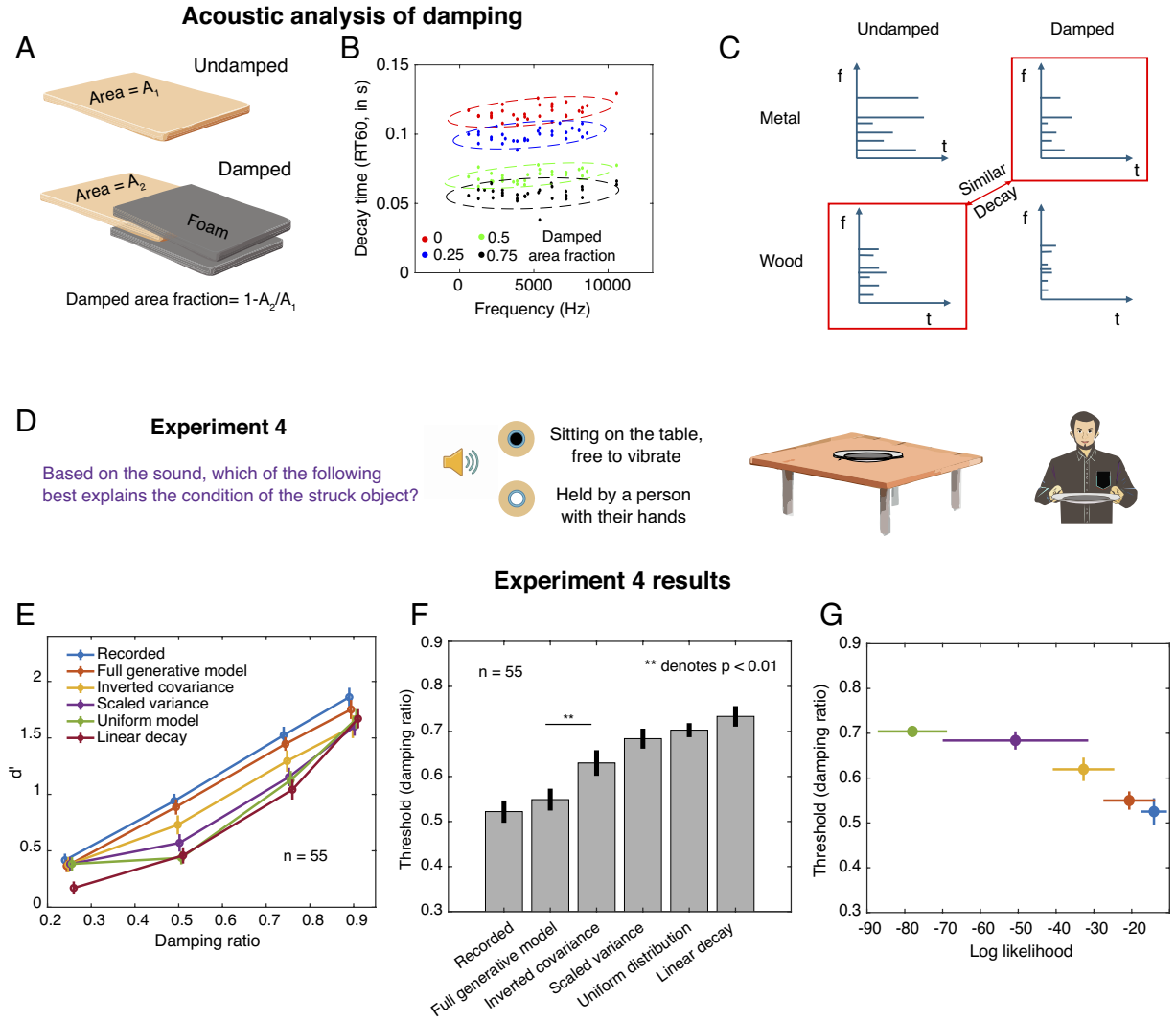



**Fig. 10.** Perception of damping. (*A*) Acoustic effects of parametrically manipulated damping. We measured impulse responses from a board that was damped to varying extents by surrounding foam, varying the amount of the board that was clamped in foam. (*B*) Frequencies and decay times of impulse response modes in each of four damping conditions. Colors correspond to the fraction of the board that was damped (see panel *A*). Decay time decreases gradually with the extent of damping, but mode frequencies are largely unaffected. Dots plot parameter estimates for individual modes. Ellipses plot 95% confidence contours to Gaussian distributions fit to the mode parameters. (*C*) Rationale for experiment: the decay times of a damped object of a resonant material (e.g., metal) might be similar to the decay times of an undamped object of a less resonant material (e.g., wood). This ill-posedness could be overcome using knowledge of the material-driven dependence of mode decay and frequency. (*D*) Experiment 4 task. Listeners heard a single impact sound and judged whether the impacted object was resting freely on a table vs. being held firmly in someone's hands. The experiment manipulated the simulated degree of damping and the extent to which the underlying impulse responses were faithful to the physics-induced regularities of real-world impacts. (*E*) Experiment 4 results: damping discrimination vs. the amount of simulated damping. Here and in F, error bars plot SEM. (*F*) Discrimination thresholds in each condition, defined as the damping ratio resulting in d-prime of 1. (*G*) Comparison of discrimination thresholds from Experiment 4 with log-likelihood of impulse responses under full generative model. The linear decay condition is omitted as the impulse response parameters were taken from the Full model, and thus do not reflect the deviant features of that model variant. Error bars on thresholds plot SEM, derived via bootstrap. Error bars on log likelihood plot SD across impulse responses.

realistic impact sounds, enabling experiments with realistic sounds without requiring them to be recorded by hand. Third, we estimated a distribution over real-world object impulse responses and could thus derive alternative (unnatural) distributions that deviated from it to varying extents, while quantifying the extent of deviation (Fig. 6*B*).

A second relevant area of prior work is that of sound synthesis. Most contemporary techniques for synthesizing the sounds of physical events are based on finite/boundary element methods (34) for deriving an object's resonances. Such methods can produce highly realistic sounds (6–8) when accurate physical descriptions (e.g., of object shape and material) are available. To synthesize large numbers of sounds in the absence of such detailed object descriptions, we instead characterized object acoustic properties in statistical terms. Our approach built on analysis-by-synthesis methods in which sound components are additively combined (35–37), on established source-filter frameworks, and on acoustical theory related to resonant modes (3). Our contribution relative to this body of prior work was to make empirical measurements from everyday objects, to show that the key predictions of classical physics largely hold for such objects (Figs. 2, 4, and 8 and *SI Appendix*, Figs. S1–S5, S12, S14, and S15), and to use these measurements to generate impulse responses via sampling.

**Limitations.** Our survey of object acoustics, though more comprehensive than prior work, relied on human reports of sound-producing objects, which are likely imperfect. Practical constraints limited sample sizes within material/size categories, and because we sampled real-world objects rather than idealized shapes, effects of shape and material were somewhat confounded. Our measurement apparatus also imposed constraints: Objects that were very small or very large could not be reliably measured,

Downloaded from https://www.pnas.org by 86.130.54.74 on November 6, 2026 from IP address 86.130.54.74.

and the apparatus itself introduced some degree of damping. We characterized objects exclusively via impulse responses, estimating stiffness from theoretical material constants. We consider this choice reasonable given that validation experiments (*Materials and Methods*) suggested the assumed coefficients were close to the actual values. Our generative model neglects sound radiation (which may matter for large objects or position-dependent listening) and assumes linearity, precluding modeling of objects that bend or break. Our experiments were also restricted to scenarios where only one impacting object radiates sound; future work could examine how well listeners can separate contributions from both objects. Our experiments show that humans can estimate one physical variable despite variation in other variables, but leave open whether humans explicitly estimate these concurrent variables. Finally, while we presume that participants' knowledge of object acoustics is implicit, we did not test the degree to which their judgments were based on unconscious versus explicit reasoning. See *Materials and Methods* for more discussion of these limitations.

**Future Directions.** Our generative model could be extended to test causal inference accounts of human perception (38). Being able to synthesize realistic sounds for physical events should enable tests of whether humans can separate physical variables from other environmental effects, e.g., reverberation (23, 26, 39, 40), spectral coloration (41), and distance (42). In addition, progress in rendering sounds of scraping, rolling, and sequences of impacts (43) should enable analogous studies of auditory perception with more extended and realistic sequences. The perception of physical events is also a promising problem with which to study audio-visual integration (44). More broadly, the general problem of perception of physical variables from sound seems likely to have broad relevance to other species. As such, it may be a good model problem for auditory neuroscience.

## Materials and Methods

We measured impulse responses from hundreds of everyday objects and built a generative model of impact sounds. This model was used to synthesize stimuli for online experiments testing perception of material, mass, and damping, analyzed using standard and permutation-based statistics. Full methods are provided in *SI Appendix, SI Materials and Methods*. All participants provided informed consent, and the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects approved all experiments.

**Data, Materials, and Software Availability.** All data and code are available via the project repository (45).

**ACKNOWLEDGMENTS.** Work supported by NSF Grant BCS-1921501, NIH fellowship F32DC013703 to J.T. and a K. Lisa Yang ICoN fellowship to V.A. We thank the McDermott Lab, Neville Hogan, and Nancy Kanwisher for comments on the paper.

Author affiliations: [a]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [c]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; [d]Department of Psychological and Brain Sciences, University of Iowa, Iowa, IA 52242; [e]Iowa Neurosciences Institute, Carver College of Medicine, University of Iowa, Iowa, IA 52242; and [f]Program in Speech and Hearing Biosciences and Technology, Harvard, Cambridge, MA 02115

1. W. W. Gaver, What in the world do we hear? An ecological approach to auditory source perception *Ecol. Psychol.* **5**, 1–29 (1993).
2. J. W. S. B. Rayleigh, *The Theory of Sound* (Macmillan, 1896).
3. P. M. Morse, K. U. Ingard, *Theoretical Acoustics* (Princeton University Press, 1968).
4. N. H. Fletcher, T. D. Rossing, *The Physics of Musical Instruments* (Springer, 2010).
5. W. W. Gaver, How do we hear in the world?: Explorations in ecological acoustics. *Ecol. Psychol.* **5**, 285–313 (1993).
6. J. F. O'Brien, C. Shen, C. M. Gatchalian, "Synthesizing sounds from rigid-body simulations" in *Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (ACM Press, San Antonio, TX, 2002), pp. 175–181.
7. N. Raghuvanshi, M. C. Lin, Physically based sound synthesis for large-scale virtual environments. *IEEE Comput. Graph. Appl.* **27**, 14–18 (2007).
8. C. Zheng, D. L. James, Toward high-quality modal contact sound. *ACM Trans. Graph.* **30**, 38 (2011).
9. D. J. Freed, Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. *J. Acoust. Soc. Am.* **87**, 311–322 (1990).
10. R. L. Klatzky, D. K. Pai, E. P. Krotkov, Perception of material from contact sounds. *Presence Teleop. Virt.* **9**, 399–410 (2000).
11. F. Avanzini, D. Rocchesso, "Controlling material properties in physical models of sounding objects" in *Proceedings of the International Computer Music Conference* (International Computer Music Association, San Francisco, CA, 2001), pp. 91–94.
12. B. L. Giordano, S. McAdams, Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *J. Acoust. Soc. Am.* **119**, 1171–1181 (2006).
13. G. Lemaitre, L. M. Heller, Auditory perception of material is fragile while action is strikingly robust. *J. Acoust. Soc. Am.* **131**, 1337–1348 (2012).
14. Z. Ren, H. Yeh, R. Klatzky, M. C. Lin, Auditory perception of geometry-invariant material properties. *IEEE Trans. Vis. Comput. Graph.* **19**, 557–566 (2013).
15. J. Hjortkjær, S. McAdams, Spectral and temporal cues for perception of material and action categories in impacted sound sources. *J. Acoust. Soc. Am.* **140**, 409–420 (2016).
16. M. Grassi, Do we hear size or sound? Balls dropped on plates. *Percept. Psychophys.* **67**, 274–284 (2005).
17. M. Grassi, M. Pastore, G. Lemaitre, Looking at the world with your ears: How do we get the size of an object from its sound? *Acta Psychol.* **143**, 96–104 (2013).
18. R. A. Lutfi, C. J. Liu, C. N. Stoelinga, Auditory discrimination of force of impact. *J. Acoust. Soc. Am.* **129**, 2104–2111 (2011).
19. S. Lakatos, S. McAdams, R. Caussé, The representation of auditory source characteristics: Simple geometric form. *Percep. Psychophys.* **59**, 1180–1190 (1997).
20. A. J. Kunkler-Peck, M. T. Turvey, Hearing shape. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 279–294 (2000).
21. J. Traer, M. Cusimano, J. H. McDermott, "A perceptually inspired generative model of rigid-body contact sounds" in *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx-19)* (Birmingham City University, Birmingham, UK, 2019).
22. J. H. McDermott, The cocktail party problem. *Curr. Biol.* **19**, R1024–R1027 (2009).
23. J. Traer, J. H. McDermott, Statistics of natural reverberation enable perceptual separation of sound and space. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E7856–E7865 (2016).
24. D. Kersten, P. Mamassian, A. Yuille, Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
25. L. H. Chen, D. G. Schweikert, Sound radiation from an arbitrary body. *J. Acoust. Soc. Am.* **35**, 1626–1632 (1963).
26. J. H. McDermott, V. Agarwal, J. Traer, Physics, ecological acoustics and the auditory system. *Curr. Biol.* **34**, R1–R8 (2024).
27. M. Aramaki, R. Kronland-Martinet, Analysis-synthesis of impact sounds by real-time dynamic filtering. *IEEE Trans. Audio Speech Lang. Process.* **14**, 695–705 (2006).
28. R. A. Lutfi, "Human sound source identification" in *Springer Handbook of Auditory Research: Auditory Perception of Sound Sources*, W. A. Yost, A. N. Popper, R. R. Fay, Eds. (Springer, 2008), vol. 29, pp. 13–42.
29. L. D. Landau, E. M. Lifshitz, *Mechanics* (Pergamon, 1960).
30. K. J. P. Woods, J. McDermott, Schema learning for the cocktail party problem. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3313–E3322 (2018).
31. M. J. McPherson, J. H. McDermott, Time-dependent discrimination advantages for harmonic sounds suggest efficient coding for memory. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 32169–32180 (2020).
32. J. Traer, S. V. Norman-Haignere, J. H. McDermott, Causal inference in environmental sound recognition. *Cognition* **214**, 104627 (2021).
33. K. J. P. Woods, M. H. Siegel, J. Traer, J. H. McDermott, Headphone screening to facilitate web-based auditory experiments. *Atten. Percept. Psychophys.* **79**, 2064–2072 (2017).
34. S. Bilbao, *Numerical Sound Synthesis: Finite Difference Schemes and Simulation in Musical Acoustics* (John Wiley & Sons Ltd, 2009).
35. J. C. Risset, D. L. Wessel, "Exploration of timbre by analysis and synthesis" in *The Psychology of Music*, D. Deutsch, Eds. (Academic Press, 1982), pp. 25–58.
36. X. Serra, *A System for Sound Analysis/Transformation/Synthesis Based on A Deterministic Plus Stochastic Decomposition* (Satnford University, 1989).
37. D. Menexopoulos, P. Pestana, J. Reiss, The state of the art in procedural audio. *J. Audio Eng. Soc.* **71**, 826–848 (2023).
38. M. Cusimano, L. Hewitt, J. H. McDermott, Listening with generative models. *Cognition* **253**, 105874 (2024).
39. T. Koumura, S. Furukawa, Context-dependent effect of reverberation on material perception from impact sound. *Sci. Rep.* **7**, 16455 (2017).
40. A. Z. Ivanov, A. J. King, B. D. Willmore, K. M. Walker, N. S. Harper, Cortical adaptation to sound reverberation. *eLife* **11**, e75090 (2022).
41. C. E. Stilp, J. M. Alexander, M. Kiefte, K. R. Kluender, Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Atten. Percept. Psychophys.* **72**, 470–480 (2010).
42. P. Zahorik, F. L. Wightman, Loudness constancy with varying sound source distance. *Nat. Neurosci.* **4**, 78–83 (2001).
43. V. Agarwal, M. Cusimano, J. Traer, J. H. McDermott, "Object-based synthesis of scraping and rolling sounds based on non-linear physical constraints" in *Proceedings of the 24th International Conference on Digital Audio Effects (DAFx-20in21)* (IEEE, Vienna, Austria, 2021), pp. 136–143.
44. T. Gerstenberg, M. H. Siegel, J. B. Tenenbaum, "What happened? Reconstructing the past from vision and sound" in *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (Cognitive Science Society, Madison, WI, 2018).
45. V. Agarwal, J. Traer, J. Schwartz, J. H. McDermott, Data from "Intuitive knowledge of object acoustics enables perceptual separation of physical variables from impact sounds." GitHub. https://github.com/mcdermottLab/impact-sound-perception. Deposited 16 October 2026.

Downloaded from https://www.pnas.org by 86.130.54.74 on November 6, 2026 from IP address 86.130.54.74.

PNAS

# Supporting Information for
Intuitive knowledge of object acoustics enables perceptual separation of physical variables from impact sounds

Vinayak Agarwal[1,2,3,+], James Traer[1,3,4,5,+], Jeremy Schwartz[1], Josh H. McDermott[1,3,6]

[1]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA
[2]Department of Mechanical Engineering, MIT, Cambridge, MA, USA
[3]McGovern Institute for Brain Research, MIT, Cambridge, MA, USA
[4]Department of Psychological and Brain Sciences, University of Iowa
[5]Iowa Neuroscience Institute, University of Iowa
[6]Program in Speech and Hearing Biosciences and Technology, Harvard, Cambridge, MA, USA
+ co-first authors

Correspondence should be addressed to J.H.M
Email: jhm@mit.edu

**This PDF file includes:**

SI Materials & Methods
Table S1. Material and size categories identified by the impact sound survey
Table S2. Material and sizes of all objects whose impulse responses were analyzed
Table S3. Physical constants used to calculate material stiffness
Table S4. Stiffness constants for each material in generative model
Figure S1. Comparison of different impulse responses from the same object
Figure S2. Mode parameters for all material and size categories
Figure S3. Mode amplitude and frequency for each material and size category
Figure S4. Mode decay time and frequency for each material and size category
Figure S5. Effect of size on average mode parameters
Figure S6. Generative parameters for noise bands of impulse responses
Figure S7. Effect of ball mass on impact sounds
Figure S8. Results of Experiments 0 and 1, separated by sound presentation
Figure S9. Results of Experiment 2, separated by sound presentation type
Figure S10. Results of Experiment 3, separated by sound presentation type
Figure S11. Results of Experiment 4, separated by sound presentation type
Table S5. Full statistical reporting for Experiments 1-4
Figure S12. Acoustic effects of everyday object damping
Figure S13. Observer model performance for damping discrimination
Figure S14. Mode amplitudes are similar across different materials
Figure S15. Validation of stiffness values and forcing function assumptions
SI References

## SI Materials & Methods

### Informed consent

All participants gave informed consent. The Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects (COUHES) granted approval for all experiments.

### Documenting the objects heard in everyday human experience

*Surveying human experience with object sounds*

We sought to measure from a representative sample of the objects people experience via contact sounds in their everyday life. We conducted two online surveys asking participants to report the impact sounds they had recently heard and to report the colliding objects involved. All experiments were conducted on Amazon's Mechanical Turk platform.

Each survey asked 100 people to report the objects involved in the most memorable recent impact sound they could recall. The first survey focused on impact sounds heard within the last hour, while the second survey expanded the timeframe to include memorable impact sounds from the entire previous day (a 24-hour period). We hoped the two surveys together would reflect the range of different impact sounds encountered (or at least noticed and remembered) in daily life. Both the surveys asked participants to list very commonly heard impact sounds, and to separately list the associated objects within several size categories.

The text of the questions for the first survey was as follows:

"This is a survey of common sounds in everyday scenes. Consider the sounds you have heard in the last hour. Which objects have you heard being struck (or tapped, scraped, dropped, etc.) in this time? Try to specify the material(s) of each struck object as best as you can in a few words. Please list at least one object for each size category. If you don't recall any specific impact sounds, list objects you might expect to hear struck in this time."

The text for the second survey was the same except "the last hour" was replaced with "the last 24 hours".

Participants were given a text box to type answers into for each of five size categories: smaller than 4in. (10cm), 4-12in. (10-30cm), 1-3ft (30-100cm), 3-10ft (1-3m), and larger than 10ft (3m). Participants could list more than one object and many did. No examples were given, so as not to bias participants. Participants choosing their own descriptive terms for objects and material.

No geographic filters were used but participants were requested to give their location. 82% (164/200) reported locations in the United States, 35 reported locations in India, 1 did not report a location. The age and sex of participants was not recorded. As no filters were used to screen participants (except a minimum age of 18) we expect the participant population to be typical of the Mechanical Turk pool when the surveys were run (January 2020).

From the responses, we tallied the number of objects of every named material in every size class. This procedure was performed manually by the authors. Material descriptions were quite consistent across respondents, with almost all mentioned objects being labelled with one (or several) of 17 material categories (summarized in Table S1). We categorized objects labelled with multiple materials as "composite" (which we designated as an $18^{th}$ material category). Many such composite objects were electronic devices (e.g., phones, laptops, lamps, coffeemakers, etc.) which were usually reported as a mixture of plastic, metal, and glass. A small number of responses listed very specific materials and were re-categorized into one of the broader and more

frequently mentioned classes (e.g., copper, brass, etc. to “metal”; plywood, cedar wood, etc. to “wood”; etc.). A similar categorization was performed for object names, in order to determine the number of unique objects in each material and size class.

The final survey categories are summarized in Table S1. Of 2802 responses, 475 unique objects and 18 material categories were identified. The distribution of responses was very strongly skewed towards a small subset of objects and materials in each size category. Only 107 unique objects accounted for over half of all the responses, and only 8 material categories accounted for 91% of all responses.

*Selection of objects to be measured*

Our survey of recorded object impulse responses aimed to record as many of these commonly identified objects as was practical over a range of materials and size categories. Some commonly reported objects could not be measured due to difficulties attaching the measurement apparatus (e.g., coins, paper, etc.). Other objects could not be measured because the amplifier and speaker setup could not broadcast enough power into the objects for a high-quality recording (e.g., concrete or brick walls, most of the floor materials, and many of the largest objects, etc.). For this reason, we did not attempt to make measurements for the largest size category. For some material categories (i.e., “composite” and rubber) our preliminary recordings showed that many objects in these categories were very strongly damped, with very short impulse responses that had no detectable modes and little variation beyond spectral shape (i.e., they all produced short sharp “clicks”). Moreover, the “composite” material class was highly variable in its constituent materials. Thus, despite the frequent mention of composite and rubber objects, we did not analyze the acoustics of these material categories. In addition, once we assessed the survey results and considered the constraints of the behavioral experiments we sought to run, we realized that two of the materials we intended to use in experiments (stone and cardboard) each did not have representation in one size category. To facilitate our eventual experiments we augmented the survey results with two stone objects in the third size category and one cardboard object in the first category. Ultimately, we obtained recordings of 497 objects accounting for 70% of the survey responses and providing multiple samples from 71% of the 51 non-empty material-size bins. We then excluded objects for which there was not at least one measured impulse response that met the SNR inclusion criterion (see below), or whose material was not one of the seven materials that we analyzed in this paper. This yielded 410 objects in total (Table S2), and a total of 1502 impulse responses that met the inclusion criterion. All analyses in the paper are restricted to these 1502 impulse responses.

*Survey limitations*

The survey we conducted was more comprehensive than previous measurements of object acoustics, but was nonetheless limited by practical constraints. We asked humans to name objects they had heard produce impact sounds with the intention of sampling from the distribution of everyday experience, but we had no way of validating whether the human reports were accurate. It is possible that humans were more likely to name some types of objects than others for reasons unrelated to the frequency with which they encounter them making sounds, which could have skewed the results. Our sample sizes in each size/material category were also constrained both by what humans reported hearing, and by the objects we could obtain and measure. We have some reason to think that the sample sizes were sufficient for our purposes, in that material discrimination of sounds synthesized from the generative model (which was based on statistics obtained from the measured impulse responses) showed similar patterns of confusions to those for recorded impulse responses (Fig. 5c&7d). However, the sample sizes could be a limitation in some contexts.

Our survey allowed us to make measurements from the sorts of objects that humans actually encounter, rather than being entirely reliant on simple shapes (bars, plates etc.) that have been the focus of prior studies. One cost of this approach is that if object shapes are not independent of material, as is likely to be the case in real-world settings, a real-world object survey will inevitably confound effects of shape and material. This was clearly the case to some extent for our measurements. For instance, some material classes did not show the expected effect of object size on mode frequency (lower frequencies for larger objects; Figs. S3-S5). This almost surely reflects differences in the object shapes that people reported encountering in different size classes for particular materials. For example, small ceramic objects are likely to be mugs, whereas large ceramic objects are likely to be tiles. In addition, because of the heterogeneity of shapes that were involved, we are unable to analyze effects of shape itself. The results are plausibly reflective of the objects that people actually encounter in life, but come at the cost of obscuring the contributions of different object variables. Our results are thus complementary to traditional approaches with simple shapes that help identify the effects of particular variables but that do not consider real-world objects.

**Table S1**. Material and size categories identified by the impact sound survey. The numbers listed indicate: No. of mentions (No. of unique objects; No. of objects that account for 50% of mentions). A small fraction of unique objects accounted for more than half of all responses. We only measured impulse responses from the four smallest size categories as our measurement method could not supply enough power to make useful measurements from the largest size category.

| | <10cm | 10-30cm | 30-100cm | 1-3m | >3m |
|---|---|---|---|---|---|
| Metal<br>655 (125; 19) | 143 (20; 2) | 128 (22; 3) | 115 (35; 8) | 146 (25; 4) | 123 (23; 2) |
| Wood<br>609 (80; 11) | 35 (5; 1) | 54 (12; 2) | 113 (22; 3) | 245 (21; 2) | 162 (20; 3) |
| Composite<br>574 (177; 21) | 57 (24; 5) | 190 (42; 4) | 193 (42; 3) | 73 (39; 5) | 61 (20; 4) |
| Plastic<br>444 (89; 14) | 140 (22; 3) | 137 (21; 4) | 107 (24; 3) | 42 (11; 2) | 18 (11; 2) |
| Glass<br>142 (34; 11) | 13 (7; 3) | 28 (4; 1) | 17 (8; 3) | 61 (8; 2) | 23 (7; 2) |
| Ceramic<br>74 (14; 7) | 12 (2; 1) | 38 (4; 2) | 12 (4; 2) | 12 (4; 2) | |
| Stone<br>54 (8; 4) | 6 (2; 1) | 4 (3; 1) | | 11 (3; 1) | 33 (3; 1) |
| Rubber<br>51 (12; 3) | 24 (6; 1) | 16 (4; 1) | 11 (2; 1) | | |
| Paper<br>47 (12; 5) | 13 (5; 2) | 27 (5; 2) | 17 (2; 1) | | |
| Concrete<br>41 (3; 1) | | | | | 41 (3; 1) |
| Cardboard<br>24 (5; 3) | | 7 (2; 1) | 14 (2; 1) | 3 (1; 1) | |
| Plaster (wall)<br>24 (2; 1) | | | | | 24 (2; 1) |
| Drywall (wall)<br>19 (2; 1) | | | | | 19 (2; 1) |
| Brick (wall/road)<br>15 (2; 1) | | | | | 15 (2; 1) |
| Foam<br>10 (4; 2) | 3 (2; 1) | 7 (2; 1) | | | |
| Asphalt (road) 8 (1;1) | | | | | 8 (1; 1) |
| Linoleum/Vinyl (floor)<br>8 (1; 1) | | | | | 8 (1; 1) |
| Laminate<br>(floor/counter)<br>3 (2; 1) | | | | | 3 (2; 1) |
| **Total<br>2802<br>(475; 107)** | 436<br>(95; 20) | 636<br>(121; 20) | 571 (141; 25) | 593 (108; 19) | 538 (97; 21) |

**Table S2**. Material and sizes of all objects whose impulse responses were analyzed.

| | <10cm | 10-30cm | 30-100cm | 1-3m | **Total (410)** |
|---|---|---|---|---|---|
| Glass | 10 | 8 | 5 | 3 | **26** |
| Ceramic | 4 | 23 | 6 | 1 | **34** |
| Metal | 8 | 36 | 3 | 5 | **52** |
| Stone | 1 | 4 | 2 | 1 | **8** |
| Wood | 18 | 30 | 33 | 28 | **109** |
| Plastic | 69 | 70 | 18 | 5 | **162** |
| Cardboard | 1 | 12 | 2 | 4 | **19** |
| **Total (410)** | **111** | **183** | **69** | **47** | |

**Object impulse response measurement**

*Object impulse response measurement apparatus*

The measurement apparatus consisted of a contact speaker that supplied vibrations to the object, contact microphones to measure the vibrations of the object in response to the speaker, and a soundcard that interfaced with the speaker and microphones and was connected to a MacBook Pro. We used two different sizes of speakers and microphones. Larger speakers and microphones have a flatter frequency response, and were used unless the object was too small to accommodate them.

Speakers: We used Dayton Audio surface thruster speakers. This type of speaker is designed to be mounted on flat surfaces and to produce sound in a wide dispersion pattern. This made it well suited to supply vibrations to an object through its surface. The speakers were either 32 mm or 64 mm in diameter.

Microphones: We used either EOOCVT (30 mm in diameter) or UXCELL (10 mm in diameter) piezo surface microphones, also designed to be mounted to a surface. Microphones were placed at 4-8 locations on flat portions of the sides or top of each object.

An 8-channel audio interface (Behringer U-PHORIA) was used to simultaneously record the response to the input signal from the microphones. Recordings were made using Audacity.

Input Signal: A noise signal consisting of the concatenation of 30 3s Golay sequences was used as the input signal. Golay sequences are complementary pairs of binary sequences (consisting of +1 and -1 elements) constructed so that their aperiodic autocorrelation functions sum to zero at all non-zero time shifts. This property enables the cancellation of measurement noise, leading to improved impulse response estimation[1].

*Object impulse response measurement procedure*

We first cleaned the exterior surfaces of the object to be measured. We then positioned the object, with the goal of stabilizing it without excessively damping its vibrational response. This was achieved with dedicated supporting structures made of rubber that are commonly used for woodworking (Rockler Work Bench Cookie Plus work grippers). The speaker and microphones were then attached to relatively flat areas of the object surface using electrical tape and plugged into the audio interface. We attached as many as 8 microphones if there was room, but sometimes had as few as 4 for smaller objects without sufficient flat surface area. The connection to the object surface was checked by a) applying a light lateral force and assessing if the speaker or microphone wiggled at all and b) making an initial recording and ensuring that there was no buzzing. Once connected, we adjusted the speaker power and recorder gain, with the goal of maximizing the signal relative to the noise floor while avoiding distortion (rattling or buzzing caused by excessive speaker power or clipping caused by excessive recorder gain). We then

broadcast the input signal and simultaneously recorded all 4-8 microphone channels using Audacity.

Once a recording was complete, we cross-correlated each channel of the recording with the input sequence to get 30 snapshots of the impulse response at each speaker- mic location pair. The impulse response measurement was the average of these 30 snapshots.

*Justification for impulse response normalization*
We opted to normalize all impulse response recordings in order to be able to arbitrarily adjust the gain on the speaker and microphones used for measurement. There were two motivations for this choice. First, to make sure every recording was adequately powered, we had to change the gain on the speaker and microphones from object to object. Specifically, being able to infer the mode parameters relied on there being "enough" power in the modes relative to the noise floor. For less resonant objects it helped to increase the gain to levels that would produce distortion in more resonant objects. Second, the distance between the surface speaker and the microphones influences the absolute level of the impulse response, but was not strongly indicative of general object properties. Speaker-microphone distances were larger for larger objects (because we sought to space the microphones out as much as possible, to maximize the representativeness of the measurements) and were also determined by idiosyncrasies of the object shape (because we could only place the speaker and microphones on flat surfaces). As a result, the absolute level of our measurements was not likely to capture anything meaningful about individual objects. However, it remains possible that there could be variation in absolute level across materials that the normalization we employed would obscure. Because absolute level is an important cue to auditory events[42], it was important to ensure that we were not discarding this information were it present. To assess whether this was the case we conducted an experiment to measure these levels in impact sounds made from different materials.

We conducted an experiment in which we recorded impact sounds from multiple objects to assess how the initial amplitude varied across materials. Because the sound amplitude is also influenced by the amplitude of the force of impact, we held this constant by fixing the incident momentum (by using a fixed pellet dropped from a fixed height above the resonating surface).

We used a Shure SM57 microphone connected to a Scarlett 6i6 audio interface to record the sounds on a MacBook Pro in a sound attenuating booth. The microphone was placed about 70 cm away from the point of impact. We chose not to use a surface microphone for this experiment because pilot experiments revealed that they exhibited distortion given the relatively high impact force used to make the recordings.

We dropped a metal pellet weighing 50 grams on different locations over an exterior surface of the object being measured. The gain on the sound interface was held constant across all recordings (set to a level that captured the maximum dynamic range across objects without introducing clipping). We estimated the initial sound amplitude to be the maximum value of the measured sound waveform. We then plotted these the means of these amplitudes for each object (computed over the 10-15 impacts we recorded per object).

We found that these amplitudes did not systematically vary with material (Fig. S14). Each material exhibited a range of amplitudes, depending on the object, but the range of amplitudes was similar for different materials. The main factor we found to influence amplitude was surface thickness, with thicker surfaces having lower initial amplitudes than thinner surfaces. This is expected given that the first vibrational wave inside the object has to travel though more material when the surface is thicker, absorbing more energy. Moreover, thin plates and beams, being more easily flexed in

one direction, can support "bending modes"[2] which create more surface displacement than similar intensity compression waves in thicker objects. Given that our analysis and experiments focused on effects of material (and overall object size), this experiment suggests that the impulse response normalization did not discard information that was critical for our purposes.

*Inclusion criteria for impulse response recordings*
Despite our attempts to select objects and recording gains that would permit high-quality impulse response measurements, in some cases we observed that the noise level in the estimated impulse response was high enough to potentially interfere with the parameter estimation. To systematically screen for such noise-corrupted measurements, we leveraged the fact that the impulse responses always decayed over time with an approximately exponential shape until they hit the noise floor. As a result, in the log-amplitude domain, the envelope of the impulse response waveform could be fit with an "elbow function" (a piecewise linear function that concatenated a sloping line segment with a horizontal line segment). We computed the amplitude envelope as the mean amplitudes for successive 5ms segments of the signal. The noise floor was reflected by the horizontal portion of the function. We estimated the noise floor for each impulse response relative to the peak amplitude as the height of the flat portion of the function relative to the y-intercept. We then discarded all impulse response measurements that had a relative noise floor higher than -30 dB.

*Limitations of impulse response measurements*
It was not possible to measure impulse responses from very small objects (because there was not enough surface area to attach the speakers and microphones to the object). We also did not attempt to comprehensively make measurements from very large objects, because our pilot experiments made it clear that our apparatus could not supply enough power to measure an impulse response that was sufficiently far above the noise floor for our purposes. This difficulty was compounded by the fact that very large objects tend to be very highly damped compared to smaller objects (evident in Fig. S2, for instance). We also note that the measurement apparatus itself inevitably influences the measurement by imposing some degree of external damping on the object being measured. The degree of damping is determined by the surface area of the microphones and speaker used for the measurement, and thus is probably not more than what would result from an object resting on a table, but we did not quantify this effect.

The choice to characterize objects exclusively by their impulse responses could also be viewed as a limitation. In particular, although an object's stiffness also influences the sound it produces on impact, we estimated stiffness from theoretical calculations based on known material-specific constants rather than measuring stiffness in every object. This choice was partly due to practical constraints. Measuring stiffness required recording impact sounds and then fitting our parametric impact sound model to the recording (see Fig. S15 for an example, and see below for details), and it was impractical to obtain clean recordings for each of the objects in the survey (this would have required transporting each of them into a sound booth, whereas the impulse response measurement could be made in the object's home environment). We view this choice as reasonable on the grounds that the impulse response is a high-dimensional quantity that contains substantially more structure (70 parameters in our parametric model) than the single scalar value captured by the stiffness coefficient. In addition, the experiments we conducted to validate the stiffness coefficients used in our generative model indicate that these assumed coefficients are close to the actual values.

**Selection of objects to include in experiments**
For the behavioral experiments described below, we drew impulse responses from a subset of the full set of measurements (all impulse responses from metal, glass, ceramic, stone, wood,

plastic, and cardboard in each of the four size categories).

**Physics-based synthesis model for object impulse responses**

*Modal synthesis of impulse response*

Building on prior work, we model object IRs as a sum of decaying sinusoidal modes and a noisy transient[3-6]:

$$h_{ir}(t) = h_{sin}(t) + h_{noise}(t)$$

where $y_{ir}(t)$ is the IR waveform and $y_{sin}(t)$ and $y_{noise}(t)$ are the decaying sinusoidal modes and the noise transient, respectively. The key assumption is that the functional form of the decay is exponential and hence can be modeled with two parameters per mode: an initial amplitude and a decay rate. See below for description of validation of this assumption. The decaying sinusoidal modes can thus be expressed as:

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

where $a_j$ is the initial amplitude (in dB), $b_j$ is the RT60 (in sec) and $f_j$ is the frequency (in Hz) of the $j^{th}$ mode respectively. For the purpose of this work, we only considered the 10 most salient sinusoidal modes for each IR recording (N = 10), both when analyzing and synthesizing IRs. The noise transient is modeled as a sum of decaying noise bands:

$$h_{noise}(t) = \sum_{i=1}^{M} 10^{\frac{\alpha_i}{20} - \frac{(60/20)}{\gamma_i}t} v_i(t)$$

where $\alpha_i$ denotes the initial amplitude (in dB re: a maximum possible value) and $\gamma_i$ denotes the RT60 (in sec) for the amplitude envelope on the $i^{th}$ noise band ($v_i(t)$). Unlike the sinusoidal modes, whose frequencies varied across impulse responses, the noise bands had fixed frequency limits, because the bands together spanned the entire audible spectrum. The noise bands were generated by filtering Gaussian white noise with a simulated bank of bandpass filters that were equally spaced on a frequency scale intended to match that of the human cochlea (the $ERB_N$-scale[7]), with the intention of making the spectral detail equally discriminable to humans across the spectrum. Specifically, we first generated a random Gaussian noise sample and then filtered it using FIR filters whose cutoffs were equally spaced on an ERB scale. We used M = 20 noise bands, as our pilot experiments indicated that this number was sufficient to account for the perceptible variations in spectral shape in everyday IRs.

**Inferring impulse response mode parameters through analysis-by-synthesis**

Because we sought to analyze distributions of impulse responses in terms of mode parameters, we first had to estimate these parameters for each measured impulse response $h_m(t)$. We used gradient-based optimization to estimate the parameters that best accounted for a measured impulse response (i.e., that minimized a loss function quantifying the difference between $h_m(t)$ and the result of the parametric function fit, $h_s(t)$).

This optimization problem is challenging because many choices of loss function exhibit local minima. We settled on a multi-scale spectrogram-based loss function that quantified the difference between $h_m(t)$ and $h_s(t)$. Due to the tradeoff between temporal and spectral resolution in spectrograms, spectrograms derived from longer time windows yield better estimates of sinusoidal mode frequencies whereas those from shorter time windows are better for estimating

temporal structure (here the RT60s of the modes and noise bands). We found that a multi-resolution spectrogram loss based on four different spectrograms (using time windows of 4096, 1024, 256, and 64 samples, respectively); the loss was the sum of the Huber loss[8] between these representations of the two signals $h_m(t)$ and $h_s(t)$ [$H_{m,n}(f,t)$ and $H_{s,n}(f,t)$ respectively]. We used the Huber loss because it mitigates problems that occur for L1 and L2 losses (L1 causes slow convergence away from zero, and L2 causes slow convergence near zero). We also added a Huber loss term between the full spectra $H_m(f)$ and $H_s(f)$ to assist the noise mode fitting:

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } a \le |\delta| \\ \delta \cdot \left(|a| - \frac{1}{2}\delta\right) & \text{otherwise} \end{cases}$$

$$L\big(h_m(t), h_a(t)\big) = L_\delta\big(H_m(f) - H_s(f)\big) + \sum_{n \in \backslash\{64,256,1024,4096\backslash\}} L_\delta\left(H_{m,n}(f,t) - H_{s,n}(f,t)\right)$$

We used $\delta$=3 in the Huber loss.

We performed gradient-based optimization in PyTorch 2.0 using differentiable audio tools contained in TorchAudio toolbox. We found that the success of the optimization was greatly aided by “good” initial guesses for each parameter. Since the parameters have physical meaning, we used physically motivated heuristics to initialize inference.

The sinusoidal mode frequencies $\{f_i\}$ were most important to initialize well because when initialized mode frequencies are sufficiently different from their true values, small changes to the mode frequencies leave the spectrogram-based reconstruction loss unchanged, preventing gradient-based optimization from getting off the ground. To initialize the mode frequencies, we used the power spectrum of $h_m(t)$ to detect possible modes, using peak picking to identify the top N candidate modes with the highest average power. We applied A-weighting to the spectrum prior to peak picking to emphasize perceptually relevant regions of the spectrum. To avoid assigning multiple peaks to the same mode, we constrained the selected peaks to be at least 100 Hz apart. To avoid explaining parts of the noise transient with sinusoidal components, we required the peaks to have a prominence ≥ 2 (to reject noise bands, that were typically wider than sinusoidal modes). Peak prominence quantifies how much a peak stands out within its local neighborhood in the signal. To determine the prominence of a peak, the minimum signal value on each side of the peak is identified within the region between the peak and the closest peak of higher amplitude. The prominence is calculated as the difference between the larger of these two signal values and the peak's height. We used the find_peaks implementation in scipy 1.14.1.

We initialized sinusoidal mode amplitudes $a_i$ by uniformly sampling from [-10 dB, -30 dB]. To initialize mode RT60s $b_i$ we first calculated the ‘broadband’ RT60 (the time it takes for the average power to dip by 60 dB) for the IR waveform, and then randomly initialized mode RT60s around the broadband RT60, uniformly sampling from [RT60-50ms, RT60+50 ms]. We found that this helped to ensure that the modes were fit correctly for both resonant and damped materials, and that sinusoidal modes were not erroneously accounted for by noise bands during optimization.

For the noise parameters $\alpha_j$, initial values were sampled from a uniform distribution according to the power ($P_j$) contained in each of the cochlear frequency bands. For each band, the initial noise amplitude($\alpha_j$) was initialized as a sample from U[$P_j$ - 20, $P_j$], with power measured in decibels. The noise RT60 $\gamma_j$ values were initialized as samples from U[0.04 s, 0.12 s].

For gradient-based optimization during inference, we used the Adam optimizer in PyTorch 2.0. We used a learning rate of $2 \times 10^{-6}$. Larger learning rates produced poor results, presumably because the loss landscape was not very smooth.

**Verifying discrete resonant modes of impulse responses (Figure 2d)**
Although there is theoretical motivation for the assumption of discrete resonant modes[9], it was important to test whether it held true for real-world objects. We evaluated the issue by measuring the fraction of the impulse response power that was accounted for by the N strongest sinusoidal modes. Specifically, for each record impulse response, we inferred the best-fitting modes for N = 2, 5, 10, 15, and 20 (fitting only the sinusoidal component of the parametric model, without the noise bands). We then synthesized an impulse response from these fitted modes and measured the ratio of the power of this synthesized impulse response to that in the recorded impulse response. As shown in Fig. 2d, the power accounted for by sinusoidal modes increases up to 10 modes, with little increase, on average, after that. This analysis supports the structural assumption of discrete modes that is built into the generative model.

**Verifying decay properties of impulse responses (Figure 2e)**
Although there is also theoretical motivation for the assumption of exponential decaying modes[9], it was again important to test whether it held true for real-world objects. To test the validity of the assumption, we performed the following procedure on each of the measured impulse responses. First, we calculated the power spectrum of the impulse response, and detected peaks using the same procedure described in the previous section (constraining peaks to be at least 100 Hz apart, and requiring them to have a peak prominence greater than 2). Second, we generated a spectrogram of the impulse response using with 512 fft points, window size of 128 and an overlap of 50% between consecutive windows to get sufficient resolution in time. Third, for each of the top 10 spectral peaks, we took the time-varying log-magnitude from the spectrogram at the peak's frequency and excerpted the portion above the noise floor (estimating the noise floor as the mean log-magnitude from the last 100 ms of the impulse response, and excerpting the portion of the prior to the first point where the log-magnitude was less than or equal to this value). Fourth, we fit polynomials of degree 0, 1, 2, and 3 to the log-magnitude vs time data for each frequency using the fit function in MATLAB (which performs least-squares minimization), and calculated the goodness of fit.

As shown in Fig. 2e, this analysis showed that there was little improvement beyond degree 1, indicating that linear functions of time provided a good fit to the mode decay on a dB scale (linear functions on a dB scale are equivalent to exponential functions on a linear scale). This verified our assumption of exponential amplitude decay with time and established this as a unifying property of object impulse responses.

**Verifying broadband onset of impulse responses (Figure 2f)**
We also verified that most impulse responses contain an initial broadband onset (that in our model is accounted for by decaying noise bands). For this analysis we measured the spectral flatness[10] of successive 20-ms-long segments of each impulse response. As shown in Fig. 2f, the spectral flatness was high in the initial segment and lower for all subsequent segments, consistent with the common presence of an initial broadband onset.

**Analysis of impulse responses from the same object (Figure S1)**
Visual inspection of impulse responses from the same object reveals a fair degree of consistency along with some variation (Fig. S1a). To quantify this consistency, we first calculated the difference in the average mode parameter values for all pairs of impulse responses within an

object, compared to the same quantity for pairs from different objects (Fig. S1b). We quantified the difference as the Euclidean distance between the mean mode parameters of two impulse responses. We measured the average distance for all pairs of impulse responses from the same object, for all pairs of impulse responses from different objects of the same size and material category, for all pairs of impulse responses from different objects of different sizes but the same material category, and for all pairs of impulse responses from different objects of the same size but different materials. This analysis revealed that mode parameters were substantially more similar for impulse responses from the same object than for impulse responses from different objects.

We then assessed the extent to which the different impulse responses from the same object share the same mode frequencies. We estimated mode frequencies for each impulse response and then calculated the number of pairs of mode frequencies that differed by 1% or less. We then made the same measurement for different objects from the same material category. We used this same measure to calculate the number of modes for each object that are only present in one of the impulse responses measured from the object ("orphan" modes). We then compared this number to that for groups of impulse responses of the same size as the within-object impulse responses, but with each impulse response from a different object (one across-object group matched to each object). This analysis showed that pairs of impulse responses from the same object tend to share a considerable number of modes (6.3 on average, SD = 0.91; Fig. S1c), and that there are very few modes that are only present in one of the impulse responses measured from the same object (Fig. S1d). Given the reversibility of linear systems, the same conclusions hold for impulse responses from different points of contact (as if we had measured impulse responses from one point but with different speaker locations).

These results confirm that real-world objects possess a modest set of resonant modes, and that any single impulse response measured from the object surface tends to include a subset of them.

**Generative model of impulse responses**

Our generative model of impulse responses sampled sinusoidal mode and noise parameters from distributions fitted to the estimated impulse response parameters for a particular material and size. For simplicity and ease of sampling we used Gaussian distributions that were truncated to limit the distribution to physically realizable values, and sampled modes independently. The distributions of sinusoidal modes thus captured the mean and variance of each of the three sinusoidal mode parameters (frequency, decay time, and starting amplitude) as well as the covariance between parameters. The distributions for noise captured the mean, variance and covariance of the decay time and starting amplitude of a frequency band.

*3D Gaussian model of sinusoidal mode parameters*

After collecting all modes for a material and a size category, we removed all modes with RT60 < 0.01s and mode frequencies < 20Hz. We then computed sample means and covariance matrices, and used these as the mean and covariance for the Gaussian for that material and size category. We then truncated the density beyond the following limits: mode frequency, [20Hz, 20000Hz]; mode amplitude, [-inf dB, 0dB]; mode RT60, [0.01s , 1s].

*2D Gaussian models of noise band parameters*

We used a similar approach to model the noise bands. The only difference was that because the noise bands in the model are fixed, we computed the means and covariances of the initial amplitude and RT60 values for each of the 20 noise bands, and had separate two-dimensional Gaussians for each band. Before estimating the means and covariances we removed noise

bands with RT60<0.005s. We then truncated the densities beyond the following limits: mode amplitude, [-inf dB, 0dB]; mode RT60, [0.005s, 1s].

*Impulse response generation*
To generate an impulse response, we first selected the material and size category, then sampled 10 sinusoidal modes (each defined by a triplet of parameters: frequency, starting amplitude, and decay time) independently from the truncated three-dimensional Gaussian, then sampled the initial amplitude and RT60 for each noise band, and then used these values in the impulse response synthesis model described above. The sampling and distribution fitting were performed in MATLAB.

**Calculation of impulse response likelihood in different synthesis conditions**
To calculate the average log-likelihood of impulse responses from a stimulus condition (described below), we used a mixture distribution obtained from a weighted sum of the distributions estimated for each material and size category, weighting each category equally. We then took each of a set of impulse responses and measured its probability under this distribution using the mode and noise band parameters, averaging the logarithm of this probability across impulse responses to obtain the results shown in Fig. 5c, 7i, and 8g. For the recorded impulse responses, we estimated the mode and noise band parameters for each recorded impulse response. For synthetic impulse responses, we sampled 10,000 sets of mode and noise band parameters from each material and size category. The probability of an impulse response was obtained as the average of the mean probabilities of each mode and the mean probability of each noise band.

**Physics-based synthesis of impact sounds from impulse responses**
*Source-filter framework*
We developed a source-filter model for synthesizing impact sounds from impulse responses, extending a previous state-of-the-art physical model[6]. The model assumes the collision interaction to be a linear system:

$$s(t) = f(t) * [h_A(t) + h_B(t)]$$

The forcing function is obtained by modeling the surfaces of colliding objects as springs. The spring model assumes that the force between a ball and the surface it impacts depends linearly upon the distance by which they are compressed when in contact. In practice, the force is subject to physical limits on the extent of deformation. We first describe the components of the forcing function that derive from the spring model, and then describe the non-linear modification we introduced to account for physical limits on deformation.

For simplicity of analysis, assume that object A is dropped onto object B, such that object B can act as an inertial frame of reference. Under this assumption, the resulting compression between A and B can be approximated as springs in series combination. The equivalent spring constant for such an interaction is given by

$$k_{eq} = \frac{k_A k_B}{k_A + k_B}$$

The resultant dynamics can be modeled by a half cycle of simple harmonic motion. This yields the forcing function to be

$f_{ideal}(t) = Ap_A sin(\omega t)$, where $t \in \left[0, \pi\sqrt{\frac{m_A}{k_{eq}}}\right]$, $A$ is a proportionality constant, $p_A$ is the momentum of object A (the product of the velocity $v$ and the mass $m_\mathrm{A}$), and $\omega = \sqrt{\frac{k_{eq}}{m_A}}$.

Because the typical duration of such a force is less than a millisecond, the primary acoustic effect of the contact force is to filter the object impulse responses, with the following transfer function:

$$F(\omega) = 2\pi v m_{\mathrm{ball}} \mathrm{sinc}\left(\frac{\omega - \kappa}{2\pi\kappa}\right)$$

where $\kappa = \sqrt{k_{\mathrm{eq}}/m_{\mathrm{A}}}$. The overall amplitude of this transfer function (and thus the resulting sound) increases with the ball mass, as expected. Moreover, the function peaks at a frequency of $\kappa/(2\pi)$. The peak frequency thus decreases for heavy balls and/or softer materials, as both of these increase the time the surface and ball are in contact, i.e., the width of the half-cycle sinusoid, resulting in a more low-pass filter (Fig. 4a&b). The formulation thus encompasses the two main acoustic cues previously associated with mass (amplitude and spectral centroid)[11,12].

The forcing function thus requires specifying the impact velocity and mass of the ball, and the stiffnesses of the ball and surface.

*Stiffness*

The forcing function introduced in the previous section depends on the equivalent spring constant, which is a function of the spring constants for the two constituent objects. These spring constants capture the stiffness of each object. Stiffness is a function of an object's material, size, and shape. Because stiffness is less straightforward to measure from an object, we calculated stiffness values from known material constants and principles of solid mechanics, verifying that these values produced sounds that were very similar to those we recorded.

Stiffness is proportional to both Young's modulus and the cross-sectional area of an object. Noting that Young's modulus varies by orders of magnitude across the materials studied in this paper, and that the object sizes we considered also varied by orders of magnitude, we made the simplifying assumption that an object's stiffness would be the same for all objects within a material and size category (because any variation between shapes is likely to be dwarfed by the variation across material and size). We then derived the stiffness for square plates of a size and material matched to each of the categories from the object survey.

Assuming a square plate of thickness h and side length a, with a central point load, and that is simply supported the deflection δ of the plate can be approximated by plate theory. The assumption of simple support entails that all four edges of the plate are free to rotate but not vertically deflect, with no non-zero moments.

The flexural rigidity D of the plate is:

$$D = \frac{Eh^3}{12(1-\nu^2)}$$

where E is Young's modulus and $\nu$ is Poisson's ratio. The central deflection for a point load F on a simply supported square plate can be approximated by:

$$\delta = \frac{Fa^2}{64D}$$

Using Hooke's law, the equivalent spring constant k can be approximated as:

$$k = \frac{F}{\delta} = \frac{64D}{a^2} = \frac{64 \cdot \frac{Eh^3}{12(1-\nu^2)}}{a^2}$$

$$k \approx \frac{16Eh^3}{3(1-\nu^2)a^2}$$

We calculated this stiffness value for side lengths of (4cm, 10cm, 30 cm and 80cm) and thicknesses of (0.5cm,1cm,1.5cm and 2cm) using material-specific values of Young's modulus and Poisson's ratio from engineering tables (https://www.engineeringtoolbox.com/), listed in Table S3.

**Table S3**. Physical constants used to calculate material stiffness

| Material | Poisson's ratio | Young's Modulus (GPa) |
|---|---|---|
| Glass | 0.2 | 70 |
| Ceramic | 0.25 | 215 |
| Metal | 0.2 | 135 |
| Stone | 0.22 | 42.5 |
| Wood | 0.3 | 11.5 |
| Plastic | 0.4 | 2.35 |
| Cardboard | 0.25 | 3.5 |

This calculation yielded the stiffness values given in Table S4. To calculate the equivalent stiffness needed for the impact forcing function, we took the stiffness values $k_A$ and $k_B$ for the materials and sizes being simulated.

**Table S4**. Stiffness constants for each material in generative model.

| Material | k (N/m) for different plate sizes (specified by side length) | | | |
|---|---|---|---|---|
| | 4 cm | 10 cm | 30cm | 80cm |
| Glass | $3.04 \times 10^7$ | $3.89 \times 10^7$ | $1.45 \times 10^7$ | $0.48 \times 10^7$ |
| Ceramic | $11.95 \times 10^7$ | $15.29 \times 10^7$ | $5.74 \times 10^7$ | $1.92 \times 10^7$ |
| Metal | $7.03 \times 10^7$ | $9.00 \times 10^7$ | $3.38 \times 10^7$ | $1.12 \times 10^7$ |
| Stone | $2.28 \times 10^7$ | $2.92 \times 10^7$ | $1.08 \times 10^7$ | $0.40 \times 10^7$ |
| Wood | $0.69 \times 10^7$ | $0.88 \times 10^7$ | $0.34 \times 10^7$ | $0.08 \times 10^7$ |
| Plastic | $0.16 \times 10^7$ | $0.21 \times 10^7$ | $0.07 \times 10^7$ | $0.024 \times 10^7$ |
| Cardboard | $0.20 \times 10^7$ | $0.25 \times 10^7$ | $0.10 \times 10^7$ | $0.032 \times 10^7$ |

In all the experiments in this paper, we assumed the impacting object to be a small metal pellet, and so set $k_B$ to the value for the smallest size category of metal. This value was the highest stiffness value across all the material-size categories. This choice had the (intentional) consequence that $k_{eq}$ was very close to $k_A$ in all cases (because $k_B$ was large). If we had instead chosen the impacting object to have lower stiffness, $k_B$ would have dominated the equivalent stiffness for some materials but not others, which seemed best to avoid for the sake of simplicity.

To help ensure that the assumptions underlying our stiffness calculation and the forcing function more generally were correct, we verified that the forcing function resulting from the assumed stiffness values closely approximated the sounds made by the everyday objects we sampled.

In a first test we recorded impact sounds made by pellets on objects of different materials and sizes and used the generative model to infer the stiffness of the impacted object. comparing the inferred value to the values we calculated (as described above). We used a set of four metal pellets of known masses (1.2, 5.9, 8.7, and 12 grams) and three different impacted objects from the impulse response survey (a small wood plank of size 4 x 8 x 2 cm, a large wood plank of size 20 x 71 x 2cm, and a large metal plank of size 15 x 60 x 1 cm). The motivation was both to test the assumption of linearity that yields a forcing function whose width depends on the stiffness-to-mass ratio, and to test whether the stiffness values needed to match actual impact sounds were close to the values we used in our model. We chose small metal pellets because their stiffness is high, and so should contribute negligibly to the equivalent stiffness of the impact. We recorded the sounds of 8 impacts of each pellet on each object. For each sound, we used analysis by synthesis with the generative model to estimate the equivalent stiffness (which was dominated by the impacted object due to the high stiffness of the metal pellets). We used an impulse response for each object that was measured during the survey described above (we picked the cleanest recording from the set that was recorded). We determined the stiffness value that minimized the reconstruction error given by the multi-resolution spectrogram loss defined in the above section on “Inferring impulse response mode parameters through analysis-by-synthesis”. We used a one-dimensional grid search for this purpose, first with a coarse grid sampling 500 stiffness values logarithmically spaced between $10^6$ N/m and $10^9$ N/m, and then with a finer search around the initial estimate $Est_i$ in the range $[Est_i\text{-}10^{0.5}, Est_i\text{+}10^{0.5}]$. This analysis yielded inferred stiffness values for each object were relatively constant across different pellet masses (Fig. S15). Moreover, the estimated stiffness values were close to the theoretical predictions that we used for stimulus generation. Both of these results help to validate the modeling assumptions.

As a second test of the overall forcing function framework, we also verified empirically that when the stiffness of the impacting object was varied, the equivalent stiffness was dominated by the softer of the two objects. Specifically, we recorded impacts made by mallets of 8 different materials when applied to a set of objects of different materials (a subset of the objects used in the impulse response survey), and then inferred the stiffness-to-mass ratio (because the mass of the mallet was not well defined, and did not matter for the purposes of the analysis because it was held constant across objects). We recorded the impact sounds using a Shure SM-50 free-field microphone. Then, using the impulse responses measured from the same objects, we inferred the object modes using the analysis-by-synthesis inference procedure described below. We then used an analogous analysis-by-synthesis inference procedure to determine the best-fitting stiffness-to-mass ratio (fixing the mode parameters). We used a 1D grid search to infer the stiffness-to-mass ratio.

This analysis confirmed that the equivalent stiffness is dominated by the softer object. Specifically, for the hardest mallets, the equivalent stiffness was determined by the object that was struck (i.e., it varied across objects). By contrast, for the softest mallet, there was little variation in equivalent stiffness across objects. This provides some additional support for the assumptions underlying the forcing function formulation that we used.

We note that for all experiments other than Experiment 1 (realism judgments), we held stiffness constant across materials to isolate the contribution of impulse response statistics. In this case (Experiments 2-5), we used a single stiffness value close to the middle of the distribution of the values in Table S4 ($k_A = 10^7$N/m).

*Nonlinear forcing function*

We found empirically that the actual forcing function produced by everyday impacts deviated from this linear model for forces that were sufficiently large (consistent with existing knowledge in material sciences on the response of real materials to impulsive load[13-15]). We accounted for this linearity limit by using a tanh function to "soft clip" the otherwise sinusoidal impact force.

$$f(t) = c_m \cdot tanh\left[Ap_{ball} \cdot sin\left(\sqrt{\frac{k_{eq}}{m_{ball}}} \cdot t\right)/c_m\right]$$

Under this model, an object is able to absorb only up to a certain amount of force that it can release in the relaxation phase. Any force above that limit is dissipated as heat and plastic deformation. The extent of the nonlinearity is governed by the incident momentum and a material- and shape-specific constant ($c_m$). For instance, thicker objects will be able to act as linear springs for higher loads than thinner counterparts of the same material. We set the constant $c_m$ by ear for each material, comparing synthesized sounds to impact sound recordings from the objects in the survey and finding a value that produced what to us sounded like a good perceptual match to a majority of the recordings for a material. The qualitative effect of the contact force noted earlier remains, with heavier masses and softer materials accentuating lower frequencies of the impulse response.

We validated the theoretical predictions for the effects of object mass using recorded impact sounds generated by dropping pellets of different masses onto object surfaces. Two examples are shown in Fig. 4d and Fig. S7a. When our simulated contact force was convolved with the measured impulse response for an object, the resulting sound was similar to the recorded audio from an actual impact in the world (Fig. S7b), helping to validate the model. The recordings also

help validate the effect of mass on amplitude: sound amplitude increased roughly linearly with mass for small pellets that stay within the linear regime of the forcing function (Fig. S7c), consistent with previous such measurements[11].

*Limitations of generative model*
Our generative model ignores the issue of sound radiation. In real objects, different points on an object's surface vibrate differently (reflected in the variation in impulse responses revealed in the analysis of Fig. S1), such that for objects of sufficient size the sound heard by an observer might depend on their position relative to the object. We neglected such issues on the assumption that they contribute modestly to everyday experience, but they likely matter in some contexts, and should be studied. The generative model also assumes linearity. This is a good assumption for rigid objects, but cannot account for objects that bend or break[16].

**Effect of damping on object impulse responses (Fig. 8b & Figure S12)**
Everyday objects are routinely damped via contact with other objects. We hypothesized that the effects of damping might be fairly simple to capture within our synthesis framework, which would facilitate the study of perceptual effects related to damping.

We first conducted two complementary experiments to measure the effect of damping on object impulse responses. In the first experiment, we recorded impulse responses from planks of four different materials (ceramic, glass, metal, and wood). To isolate the planks from their surroundings, they were placed on foam cones. Damping was then manipulated with weighted foam placed on both sides of the planks (Fig. 8a), varying the proportion of the plank that was encased in the foam (10%, 25%, 50%, and 75%). Impulse responses were measured with the apparatus described above, with the surface microphones and speaker placed on the exposed portion of the plank. Mode parameters were estimated for each measured impulse response and then analyzed for each material and damping condition.

As shown in Fig. 8b, the measured impulse responses revealed that the effect of damping was primarily confined to the mode decay rates: RT60 values decreased linearly as the damped area increased. By contrast, mode frequencies and amplitudes remained relatively unchanged for different damping levels.

To assess the effects of damping in a commonly occurring everyday scenario, we conducted a second experiment comparing impulse responses of objects held by a human to those of the same objects resting freely on a table. We selected comparatively resonant objects made of ceramic, glass, and wood (cups, plates, and bowls). In each of the two recording conditions, we placed a surface microphone on the object at 2-3 different locations. The object was then struck using the flick of a finger, and the resulting impact sound was recorded. We then ran a mode inference algorithm on the recorded sounds, assuming a linear forcing function with a stiffness of $10^8$ N/m and a mass of 0.1kg. Finally, we analyzed the inferred mode parameters across the two damping conditions for each object.

The mode frequencies and initial amplitudes again remained relatively unchanged in this second scenario when an object is damped with a human hand, whereas the mode RT60s change substantially (Fig. S12).

Taken together the results of these two experiments suggest that damping can be simulated by scaling the RT60s of an impulse response. For our experiments on the perception of damping, we computed the extent of RT60 variation for each material type by calculating the ratio of the clamped and unclamped RT60s from Experiment 2. We then defined a "damping ratio" to be

$$\text{dampingratio} = \frac{RT60_{\text{unclamped}} - RT60_{\text{clamped}}}{RT60_{\text{unclamped}}}$$

and varied this ratio from 0.1 to 0.9 to simulate different levels of damping while keeping the mode frequencies and initial amplitudes constant.

**General procedure for psychophysical experiments**

To obtain large sample sizes, we ran all experiments online. Extensive research conducted in our laboratory has consistently demonstrated that online data can match the quality of data collected in traditional laboratory settings[17-23] provided measures are taken to ensure standardized sound presentation, encourage participant compliance, and promote active engagement in the tasks.

We used Prolific to recruit participants. To participate, Prolific participants had to reside in Canada, the United Kingdom, or the United States, have an approval rating exceeding 95%, and be fluent in English. Target sample sizes were determined from power analyses using pilot versions of each of the experiments (see below for details). In practice, we typically slightly exceeded the target sample sizes due to the constraints of running online participants in batches.

To ensure standardized sound presentation and improve listening conditions, participants were asked to conduct the experiment in a quiet environment, minimizing external sounds. They were then instructed to adjust the computer volume to a comfortable level while listening to a calibration noise signal. The calibration was a pink noise signal peak-normalized to 0.5, and was used a reference to which stimulus levels were scaled. Each experiment then commenced with a "headphone check" to verify headphone use[24], promoting consistent sound presentation and reducing background noise. If participants passed this check, they proceeded to the main experiment. No feedback was given in any of the main experiments.

After the main experiment, participants completed a 20-trial pitch discrimination task, the goal of which was to screen for attentiveness and task compliance. Each trial presented two harmonic complex tones whose fundamental frequencies differed by one semitone, and participants judged which tone was higher. Participants were excluded from analysis of the main experiment if their performance on the pitch discrimination task was worse than 0.85.

**Stimulus conditions for behavioral experiments**

Here we describe the stimulus conditions used in the experiments. Each experiment used a subset of these conditions. All stimuli were synthesized using the impact sound model described above, and differed in the impulse responses used for synthesis (and for Experiment 3, they also differed in the mass values used in the forcing function).

Recorded impulse responses: Mode parameters were extracted from impulse response recordings from the survey and were used to resynthesize the impulse response. The resynthesis served to remove measurement noise, which otherwise would have differentiated stimuli generated from measured impulse responses from those from impulse responses sampled from the generative model.

Full generative model: The impulse response modes were sampled from the full generative model (i.e., using the fitted Gaussian distributions for a material and size category).

Inverted covariance: Impulse response modes were sampled from Gaussian distributions that had opposite signed off-diagonal covariance values compared to the distributions in the full

generative model (that were determined by fitting to the survey data). The noise modes were sampled by reversing the assignment of the 2D Gaussians to a reversed set of ERB bands (the parameters for the first subband were assigned to the last subband, the parameters for the second subband were assigned to the second-to-last subband, and so on).

Inverted covariance with scaled variance ("Scaled variance" in Fig. 6 and elsewhere): Impulse response modes were sampled from Gaussian distributions that had opposite signed off-diagonal covariance values compared to the distributions in the full generative model (that were determined by fitting to the survey data). In addition, the diagonal values in the covariance matrix were scaled by a factor of 5. The noise modes were sampled by reversing the assignment of the 2D Gaussians to a reversed set of ERB bands (the parameters for the first subband were assigned to the last subband, the parameters for the second subband were assigned to the second-to-last subband, and so on). The diagonal values of the individual covariance matrices for the noise modes were also scaled by a factor of 5.

Uniform distribution: Impulse responses modes, and noise band parameters, were sampled from uniform distributions fit to the survey data for each material and size category. The limits of the uniform distributions were set using the maximum and minimum values of the mode or noise band parameters each material and size category.

Inverted distribution: Impulse response modes, and noise band parameters, were sampled from distributions that were inverted about the maximum values for each mode parameter that was measured in the IR survey. This was achieved by sampling mode parameters from the Full generative model and then subtracting the parameter value from the maximum measured value for frequency, amplitude, or RT60. This shifts the mean and changes the sign of off-diagonal covariance elements.

Two modes, Gaussian: Impulse responses were synthesized from the full generative model, but using only 2 sinusoidal modes instead of 10. The noise component of the impulse response was generated as in the full generative model.

Linear decay: Impulse responses are synthesized from the full generative model but with linear decay in place of exponential decay. Specifically, the mode and noise band amplitude decayed linearly over time, with a linear function that dropped 60dB over the RT60 duration.

**Experiment 0: Material discrimination of recorded and synthetic impact sounds**

We validated the impact sound synthesis with an experiment measuring material identification using recorded and synthetic sounds. There were two reasons for using a material recognition task. First, material recognition yields a relatively high-dimensional result (a confusion matrix), providing a strong test of whether synthetic sounds capture the perceptual effects associated with real-world sounds. Second, material recognition seemed more likely to be robust to uninteresting differences between recorded and synthetic sounds, such as background noise or reverberation that would inevitably be present to some extent in recorded but not synthetic sounds.

Procedure: On each trial, a participant heard a single impact sound triggered by the click of a button on the trial page. Participants judged the material of the impacted object (responses selected from 7 options). The experiment consisted of 168 trials (four trials for each combination of the 3 stimulus conditions, 7 materials, and 2 size categories).

Stimuli: Stimuli were drawn from one of three conditions. In the first condition, stimuli consisted of recordings of impact sounds. The other two conditions were the Recorded impulse response

and Full generative model conditions described above (i.e., with impulse responses specific to those conditions combined with a forcing function as per the impact sound generative model). The latter two conditions used material- and size-specific values of stiffness, such that both the impulse response and forcing function contained information about material (as in real-world sounds).

The sounds for the recorded impact sound condition were recorded in a sound booth using a Shure SM57 microphone connected to a Scarlett 6i6 audio interface. The microphone was placed about 70 cm away from the point of impact. The objects were placed on foam supports (to minimize damping), which in turn were placed on a table. To generate impact sounds we dropped a metal pellet weighing 50 grams on each object, catching the pellet by hand after the first bounce. The objects were a mix of objects from the impulse response survey and substitutes that we were able to procure at the time of the recording. We drew stimuli from two size categories (10-30 and 30-100cm). The rationale for this choice was that these two sizes enabled high quality audio recordings (larger objects were often impractical to transport into our sound booth, and smaller objects often did not produce much sound with the 50-gram pellet). We recorded 10-20 impacts per object, extracted each impact, and pooled all the recorded impacts for a material and size category. We then randomly sampled from these to obtain experimental stimuli.

Each participant heard one example of the 7 material classes (metal, glass, ceramic, stone, wood, plastic, and cardboard) and 2 size categories for each condition. Stimuli for the Recorded impulse response condition were resynthesized from a recorded impulse response randomly drawn from the set of survey recordings in the desired material and size category. All sounds were synthesized using a mass of 0.05 kg, material-specific stiffness values given in Table S4, and an impact velocity of 2m/s. Each stimulus was peak-normalized to 0.5.
Power analysis to determine sample size: Based on the same reasoning used for Experiment 2 (which used the same task), we targeted a sample size of 50.

Participants: 148 participants were recruited through Prolific. Out of these, 97 participants were excluded either because they failed the headphone check, had self-reported hearing loss, did not complete the full experiment, or did not meet the performance inclusion criterion for the pitch discrimination screen. As a result, a total of 51 participants were included in the data analysis. Of these participants, 27 identified as female, 23 as male, and 1 as nonbinary. The average age of the participants was 36.7 (SD = 10.5). All participants were unique to this experiment.

Statistics and data analysis: To compute the confusion matrices, we pooled data across the entire participant pool for each test condition. The cell in the $i^{th}$ row and $j^{th}$ column of the confusion matrix was the proportion of trials in which the stimulus was generated from material *i* but was labeled as material *j*. We compared the confusion matrix for the two synthetic conditions to that for the recorded impact sound condition with a correlation coefficient.

**Experiment 1: Realism discrimination**
Procedure: On each trial, a participant heard two impact sounds in succession, separated by 0.25 seconds. One of the sounds was synthesized using a recorded impulse response as described in the previous section, while the other was from one of six synthetic conditions. The two sounds were presented in random order. Participants judged whether the first or the second sound came from a real object. The start of each trial was triggered by the click of a button. The experiment consisted of 196 trials (7 synthesis conditions x 7 materials x 4 sizes, presented in random order).

Stimuli: One of the two sounds on a trial was resynthesized from a recorded impulse response that was randomly drawn from the set of survey recordings used for experiments. The other sound

was synthesized from one of six conditions: Full generative model, Uniform distribution, Two modes Gaussian, Two modes uniform, Linear decay Gaussian, and Linear decay uniform. We omitted the No covariance and Inverted covariance conditions as it was clear from pilot experiments that although they altered the nature of the sampled sounds, they did not sound unrealistic (for instance, they might cause one material to sound like another, but not to sound like something that could not have occurred in the world). All of the conditions were synthesized using a mass of 0.1kg. The stiffness values were taken from the theoretical predictions given in the above section on "Stiffness". The impact velocities were set to 2m/s to keep the forcing function in the (approximately) linear regime. Each stimulus was peak-normalized to 0.5, to help ensure that all stimuli were audible. Each participant heard each material-size combination once in each synthesis condition. There were seven materials (metal, glass, ceramic, stone, wood, plastic, and cardboard) and four size categories (all four from which we had recorded impulse responses).

Power analysis to determine sample size: We sought to be well powered to detect differences between conditions using a paired t-test. Using data from a pilot experiment, we calculated the effect size for the difference between the proportion correct for the Full model and Two modes conditions, as this was a modest effect in the pilot experiment that seemed on par with the effects we would hope to detect in the main experiment. The pilot experiment was similar to the main experiment but had a different set of conditions that partially overlapped with the main experiment. The smallest effect of interest in the pilot experiment had a difference in means of 0.084, with a standard deviation of the difference between conditions of 0.24. A sample size of 67 was sufficient to achieve power of 0.8 with a significance threshold of 0.05. Here and elsewhere, we slightly exceeded this target because participants were run in batches.

Participants: 287 participants were recruited through Prolific. Out of these, 189 participants were excluded either because they failed the headphone check, had self-reported hearing loss, did not complete the full experiment, or did not meet the performance inclusion criterion for the pitch discrimination screen. As a result, a total of 87 participants were included in the data analysis. Of these participants, 45 identified as female, 40 as male, and 1 as nonbinary. The average age of the participants was 37.4 (SD = 11.9). All participants were unique to this experiment.

Statistics and data analysis: We calculated the performance of the participants as a fraction of the trials they were correct on, per condition. Error bars plot the standard error of the mean across participants. Differences between conditions were evaluated with t-tests, with effect sizes quantified as Cohen's d. For brevity we provide statistics for only the weakest comparison in the main text (the other differences with the full generative model condition were highly significant and had even larger effect sizes). See Table S5 for exhaustive reporting of statistics.

**Experiment 2: Material discrimination**

Procedure: On each trial, a participant heard a single impact sound triggered by the click of a button on the trial page. Participants judged which of 7 materials the impacted object was made of. The experiment consisted of 224 trials (one trial for each combination of the 8 stimulus conditions, 7 materials, and 4 size categories).

Stimuli: Stimuli were drawn from each of the eight synthesis conditions described above (Recorded impulse responses, Full generative model, Inverted covariances, Uniform distribution, Two modes Gaussian, Two modes uniform, Linear decay Gaussian, and Linear decay uniform). Each participant heard one example of the 7 material classes (metal, glass, ceramic, stone, wood, plastic, and cardboard) and 4 size categories for each condition. Stimuli for the Recorded impulse response condition were resynthesized from a recorded impulse response randomly drawn from

the set of survey recordings in the desired material and size category. All sounds were synthesized using a mass of 0.5 kg, a stiffness of 10^7 N/m, and an impact velocity of 2m/s. Each stimulus was peak-normalized to 0.5.
Power analysis to determine sample size: We sought to be well powered to detect a difference between the correlations for two conditions using a permutation test. Using pilot data, we computed power empirically as a function of sample size. The pilot experiment was similar to the main experiment but had a different (partially overlapping) set of conditions. We measured the power to detect the smallest difference in correlations that was of interest (the difference between the correlation between the Recorded and Full generative model conditions and the correlation between the Recorded and Inverted covariance condition) by drawing bootstrap samples of participants of different sizes, running a permutation test on each bootstrap sample, and then plotting the proportion of significant tests ($p<.05$) vs. sample size. This procedure indicated that a sample size of 50 would be sufficient to yield a power of 0.8.

Participants: 282 participants were recruited through Prolific. Out of these, 225 participants were excluded either because they failed the headphone check, had self-reported hearing loss, did not complete the full experiment, or did not meet the performance inclusion criterion for the pitch discrimination screen. As a result, a total of 57 participants were included in the data analysis. Of these participants, 30 identified as female, 26 as male, and 0 as nonbinary. The average age of the participants was 38.3 (SD = 12.1). All participants were unique to this experiment.

Statistics and data analysis: Because the confusion matrix provides a rich perceptual signature with which to probe human sensitivity to the acoustic structure of the model, the analysis compared material confusion matrices across conditions. To compute the confusion matrices, we pooled data across the entire participant pool for each test condition. The cell in the $i^{th}$ row and $j^{th}$ column of the confusion matrix was the proportion of trials in which the stimulus was generated from material *i* but was labeled as material *j*. We then compared the confusion matrix for each synthetic condition to the resynthesized impulse response condition (as the latter was our best estimate of the ground truth benchmark for human material discrimination performance). We quantified the similarity between confusion matrices with a correlation coefficient. The error bars on these correlations (Fig. 6c) plot the standard error of the mean, derived via bootstrap (10,000 samples). We evaluated the statistical significance of differences in correlations using permutation tests (i.e., evaluating the probability of the observed difference under the distribution of differences in correlations measured when condition labels were permuted within participants for the pair of conditions being compared). Effect sizes were quantified as the difference in correlation. For brevity, in the main text we provide the minimum p value and effect size rather than listing them for all conditions. See Table S5 for exhaustive reporting of statistics.

**Experiments 3a, 3b, & 3c: Mass discrimination**
We chose a discrimination task rather than an estimation task with a single impact sound because it was not obvious that typical human participants would be able to express well-calibrated absolute mass judgments[11].

Procedure: On each trial, a participant heard two impact sounds in succession, separated by 0.25 seconds. Participants were told that the sounds were produced by one object dropped on a surface, and judged whether the first or the second sound came from a heavier dropped object. The start of each trial was triggered by the click of a button. The experiment consisted of 140 trials. For Experiment 3a, the 140 trials consisted of 7 synthesis conditions x 5 mass ratios x 2 material classes (hard and soft) x 2 repetitions. For Experiments 3b and 3c, the 140 trials consisted of 7 synthesis conditions x 5 mass ratios x 4 material combinations.

Stimuli: Stimuli were drawn from seven of the synthesis conditions described above (Recorded impulse responses, Full generative model, Inverted covariances, Inverted covariance with scaled variance, Uniform distribution, Inverted distribution, and Linear decay). We omitted the Two modes condition because it seemed plausible that the smaller number of modes would produce impaired performance under any hypothesis. These seven stimulus conditions were fully crossed with five mass ratio conditions, with ratios of 1.8, 3.16, 10, 31.6, and 100. The two sounds on a trial were each generated under the same stimulus condition, with the specified mass ratio, and with the smaller of the two masses being 100 grams. All sounds were synthesized using a stiffness of 10^7 N/m and an impact velocity of 2m/s. For Experiments 3a and 3b, the sounds for the heaviest masses were peak-normalized to 0.99 (i.e., about 6 dB above the calibration noise), and all other stimuli were scaled relative to this value.

For Experiment 3a, each trial was assigned to one of two material classes: hard (metal, glass, ceramic and stone) and soft (wood and plastic). The impulse response used for the two sounds on a trial was randomly assigned to one of the materials in the class, and then the impulse response was sampled from the appropriate generative distribution.

For Experiment 3b, each trial was assigned to one of four material-combination conditions that dictated the materials from which the surfaces were sampled. The conditions drew from two groups of materials: hard (metal, glass, ceramic and stone) and soft (wood and plastic). The first condition sampled the surface materials for both sounds of a trial from the hard category. The second condition sampled both surface materials from the soft category. In both cases the two surface materials on a trial were constrained to be different, and the heavier and lighter objects were randomly assigned to the two materials. The third and fourth conditions sampled one surface material from the hard group and one from the soft group. In the third condition, the heavier object was assigned to the hard surface, and in the fourth condition, the heavier object was assigned to the soft surface.

Stimuli for Experiment 3c were identical to Experiment 3b except that the peak amplitude of each sound was set to 0.5, such that the two sounds on each trial were level-equated.

For all three experiments, all sounds were synthesized for the third size category (30-100cm), as this was appropriate for the moderately sized surface on which the objects were supposed to be dropped. Stimuli for the Recorded impulse response condition were resynthesized from a recorded impulse response randomly drawn from the set of survey recordings for the desired material and size category.
Power analysis to determine sample size: We determined a target sample size for Experiment 3b, as this was the most scientifically critical of the three experiments. We sought to be well powered to detect a difference in the thresholds for two conditions using a permutation test. Using pilot data, we computed power empirically as a function of sample size using a procedure like that for Experiment 2. The pilot experiment was similar to the main experiment but had a different (partially overlapping) set of conditions. We measured the power to detect the smallest difference in thresholds that was of interest (the difference between thresholds for the Recorded and Inverted Covariance conditions) by drawing bootstrap samples of participants of different sizes, running a permutation test on each bootstrap sample, and then plotting the proportion of significant tests ($p<.05$) vs. sample size. By extrapolating to larger sample sizes, we estimated that a sample size of 100 would be sufficient to yield a power of 0.8. We targeted the same sample size for Experiment 3a, and targeted a somewhat smaller sample size for Experiment 3c once it became clear that the effects in Experiment 3b were fairly large.

Participants:
Experiment 3a. 315 participants were recruited through Prolific. Out of these, 214 participants were excluded either because they failed the headphone check, had self-reported hearing loss, did not complete the full experiment, or did not meet the performance inclusion criterion for the pitch discrimination screen. As a result, a total of 101 participants were included in the data analysis. Of these participants, 44 identified as female, 53 as male, 2 as nonbinary, and 2 declined to answer. The average age of the participants was 35.1 (SD = 10.4). All participants were unique to this experiment.

Experiment 3b. 393 participants were recruited through Prolific. Out of these, 275 participants were excluded either because they failed the headphone check, had self-reported hearing loss, did not complete the full experiment, or did not meet the performance inclusion criterion for the pitch discrimination screen. As a result, a total of 118 participants were included in the data analysis. Of these participants, 52 identified as female, 62 as male, 3 as nonbinary, and 1 declined to answer. The average age of the participants was 36.2 (SD = 11.4). All participants were unique to this experiment.

Experiment 3c. 208 participants were recruited through Prolific. Out of these, 148 participants were excluded either because they failed the headphone check, had self-reported hearing loss, did not complete the full experiment, or did not meet the performance inclusion criterion for the pitch discrimination screen. As a result, a total of 60 participants were included in the data analysis. Of these participants, 30 identified as female, 27 as male, 1 as nonbinary, and 2 declined to answer. The average age of the participants was 38.7 (SD = 12.4). All participants were unique to this experiment. To facilitate an analysis of the results split by sound presentation (earphones vs. headphones) requested by a reviewer, we subsequently recruited an additional 24 participants (12 male and 12 female). The average age of these additional participants was 43.9 (SD=10.1). These additional participants were included only in the analysis shown in Figure S10.

Statistics and data analysis: We computed proportion correct for each mass ratio and synthesis condition for each participant (pooling across materials). We plotted the results as psychometric functions for each stimulus condition. The error bars plot standard error of the mean across participants. To summarize the performance in each synthesis condition, we estimated thresholds from the psychometric functions as the mass ratio producing 80% correct. We estimated thresholds by fitting a piecewise cubic spline to the psychometric function for each condition and calculating the mass ratio at which the proportion correct is 80%. For the error bars, we computed the SEM from bootstrapped samples of the participant pool. We evaluated the statistical significance of differences in the thresholds for two conditions using a permutation test (i.e., evaluating the probability of the observed difference under the distribution of threshold differences measured when condition labels were permuted within participants for the pair of conditions being compared). Effect sizes were quantified as the ratio between thresholds. For brevity, in the main text we provide the minimum p value and effect size rather than listing them for all conditions. See Table S5 for exhaustive reporting of statistics.

Cue discriminator models: We tested two cue discriminator models on this experiment. The first model made the mass discrimination judgment by comparing the spectral centroid of the two sounds on each trial. The spectral centroid was computed by evaluating the mean of all frequencies in the signal weighted by the linear sound amplitude at that frequency. The second model computed an estimate of amplitude using an integration window of 40 ms (applied to the first 40 ms segment of each sound, to approximate a window of integration applied to an impulsive sound). The models compared the centroid or amplitude values for the two sounds, assigning a higher mass to the sound with a lower centroid or higher amplitude. Prior to this comparison,

internal noise was added to the centroid or amplitude values, with the noise level set so as to approximately match the overall performance of the human participants on Experiment 3a (the same noise level was used for all experimental conditions).

**Experiment 4: Perception of damping**

Procedure: On each trial, a participant listened to a single impact sound, triggered by the click of a button. Participants judged which of the following two scenarios best explains the condition of the struck object: sitting on a table, free to vibrate; or held by a person with their hands. The experiment consisted of 300 trials (6 synthesis conditions x 5 damping conditions x 5 materials x 2 object sizes).

Stimuli: Stimuli were drawn from six conditions (Recorded impulse responses, Full generative model, Inverted covariances, Inverted covariance with scaled variance, Uniform distribution, and Linear decay). We omitted the Two modes condition because it seemed plausible that the smaller number of modes would produce impaired performance under any hypothesis. The Inverted distribution condition was omitted because we found the observer model based on RT60 performed better in this condition compared to the others, which made it non-diagnostic of our hypothesis (Fig. S13). These six stimulus conditions were fully crossed with five damping conditions. The five damping conditions were generated by scaling the mean RT60 values of the marginal distributions for mode RT60s by different proportions: 1 (undamped), 0.75, 0.5, 0.25, or 0.1. The damping ratios plotted in Fig. 10e&f is one minus this proportion. Stimuli were drawn from five materials (glass, ceramic, metal stone, and wood) and the two smallest size categories. Each material and size was used once in each combination of synthesis and damping condition. Stimuli for the Recorded impulse response condition were resynthesized from a recorded impulse response randomly drawn from the set survey recordings from the specified material and size category. All sounds were synthesized using a mass of 0.5 kg, a stiffness of 10^7 N/m, and an impact velocity of 2m/s. Each stimulus was peak-normalized to 0.5.

Power analysis to determine sample size: We sought to be well powered to detect a difference in the thresholds for two conditions using a permutation test. The procedure was identical to that for Experiment 3. This procedure indicated that a sample size of 50 would be sufficient to yield a power of 0.8.

Participants: 153 participants were recruited through Prolific. Out of these, 98 participants were excluded either because they failed the headphone check, had self-reported hearing loss, did not complete the full experiment, or did not meet the performance inclusion criterion for the pitch discrimination screen. As a result, a total of 55 participants were included in the data analysis. Of these participants, 30 identified as female, 22 as male, and 3 as nonbinary. The average age of the participants was 37.4 (SD = 11.9). All participants were unique to this experiment. To facilitate an analysis of the results split by sound presentation (earphones vs. headphones) requested by a reviewer, we subsequently recruited an additional 139 participants (58 male, 73 female, 8 nonbinary). The average age of these additional participants was 39.0 (SD=11.1). These additional participants were included only in the analysis shown in Figure S11.

Statistics and data analysis: We pooled the responses for each participant across materials. We then computed a false alarm rate for each synthesis condition from the undamped condition, and a hit rate for each combination of synthesis and non-zero damping ratio. We set values of 0 or 1 to 1/24 and 23/24, respectively, and then computed d’ for each participant. We plotted the results as psychometric functions for each synthesis condition. The error bars plot standard error of the mean across participants. To summarize the performance in each synthesis condition, we estimated thresholds from the psychometric functions. We estimated thresholds by fitting a

piecewise cubic spline to the psychometric function for each condition and calculating the damping ratio at which d-prime was equal to 1. For the error bars, we computed the SEM from bootstrapped samples of the participant pool. We evaluated the statistical significance of differences in the thresholds for two conditions using a permutation test, as in Experiment 3. Effect sizes were quantified as the ratio between thresholds. For brevity, in the main text we provide the minimum p value and effect size rather than listing them for all conditions. See Table S5 for exhaustive reporting of statistics.

Cue discriminator model: To avoid issues associated with having to pick a criterion, we ran the model on a 2I-2AFC version of the experiment. On each simulated trial, the damped and undamped stimuli were synthesized from impulse responses sampled independently from the same size category (chosen randomly on each trial). Material was sampled independently for the two stimuli. The model compared the average mode RT60 of a damped and undamped sound, and picked the sound with a shorter decay time as the damped stimulus. The average mode RT60 was computed from the decay times used to synthesize the impulse response.

**Author Contributions**

J.A.T. and J.H.M. designed the impulse response measurement survey. J.A.T. conducted the human survey of sound-making objects. J.A.T. and J.S. refined the impulse response measurement method and made most of the measurements. J.A.T. designed the impact sound generative model. V.A. refined the generative model by introducing the nonlinear forcing function. V.A. conceived and implemented the gradient-based procedure for estimating the parameters of recorded impulse responses. V.A. conducted the experiments validating the assumptions in the generative model about the effects of stiffness and mass. J.A.T. conducted pilot analyses of the impulse response recordings. V.A. conducted the final version of the analysis of impulse response recordings. V.A. and J.H.M. designed and piloted Experiment 0 and made the impact recordings. J.A.T., V.A. and J.H.M. designed Experiments 1-3. J.A.T. conducted pilot versions of Experiments 1-3. V.A. conducted the final versions of Experiments 0-3, and the control experiments for Experiment 3. V.A. and J.H.M. designed Experiment 4 and the experiments measuring the acoustic effects of damping. V.A. conducted Experiment 4 and the experiments measuring the acoustic effects of damping. V.A. analyzed all behavioral experiments. J.H.M. and V.A. drafted the initial version of the manuscript, with input from J.A.T. V.A. made the figures, with guidance from J.H.M. V.A., J.A.T., and J.H.M. edited the manuscript. J.H.M. revised the manuscript and figures for the final publication.

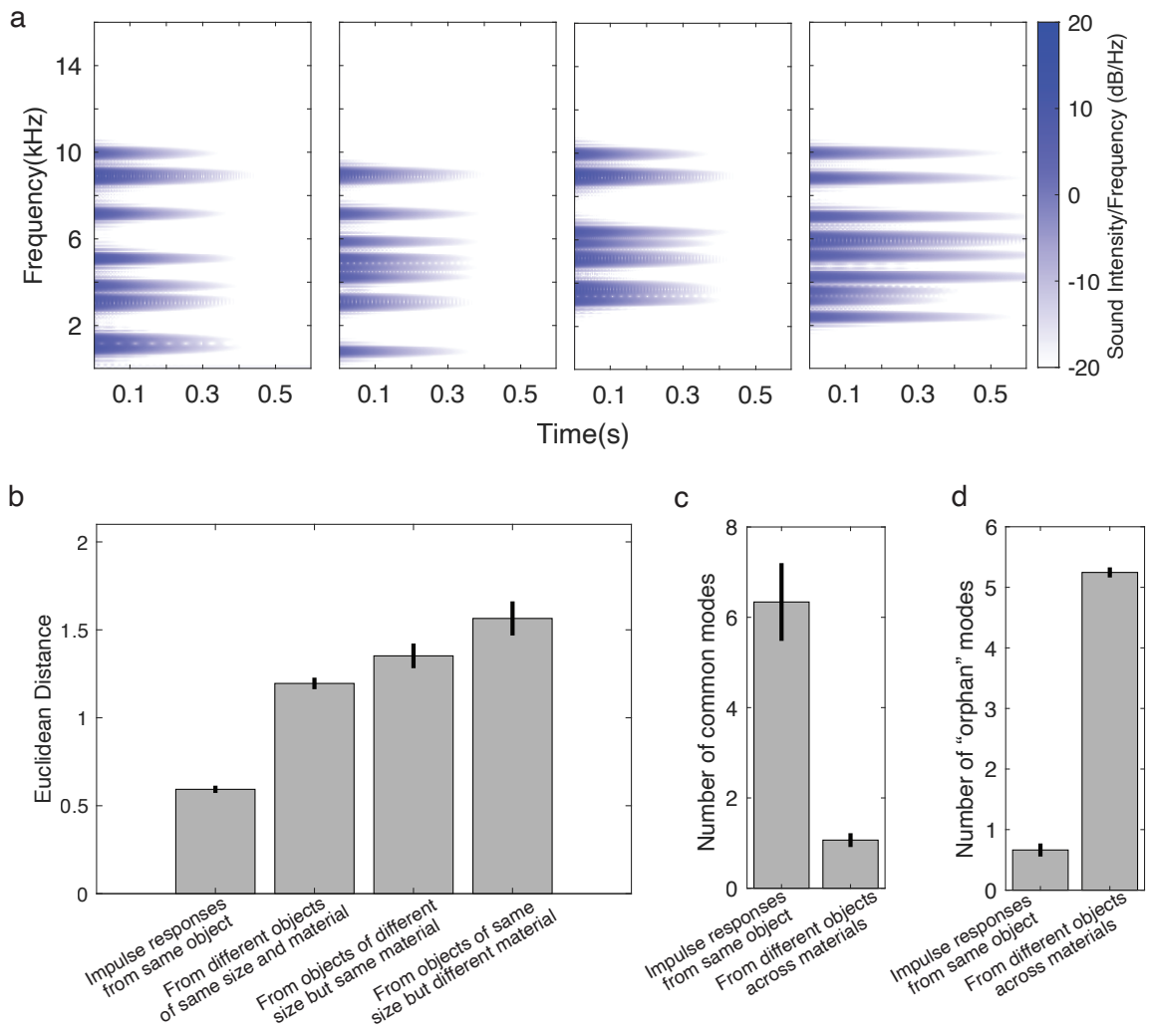



**Figure S1**. Comparison of different impulse responses from the same object. a. Spectrograms of different impulse responses measured from the same object. Visual inspection suggests that the different impulse responses share mode frequencies. b. Difference in the average mode parameter values for all pairs of impulse responses within an object, compared to the same quantity for pairs from different objects. We quantified the difference as the Euclidean distance between the mean mode parameters of two impulse responses. Graph plots the average of this distance for all pairs of impulse responses from the same object, for all pairs of impulse responses from different objects of the same size and material, for all pairs of impulse responses from different objects of different sizes but the same material, and for all pairs of impulse responses from different objects of the same size but different materials. c. Analysis of the extent to which different impulse responses from the same object share the same mode frequencies. We estimated mode frequencies for each impulse response and then calculated the number of pairs of mode frequencies that differed by 1% or less. By this measure, pairs of impulse responses from the same object tend to share a considerable number of modes (6.3 on average, SD = 0.91), whereas impulse responses from different objects do not (1.04 on average, SD = 0.14). We used this same measure to calculate the number of modes for each object that are only present in one of the impulse responses measured from the object. There were very few such modes (0.66 on average, SD = 0.11). By comparison, "orphan" modes were relatively common from groups of impulse responses drawn from different objects (5.14 on average, SD = 0.080).

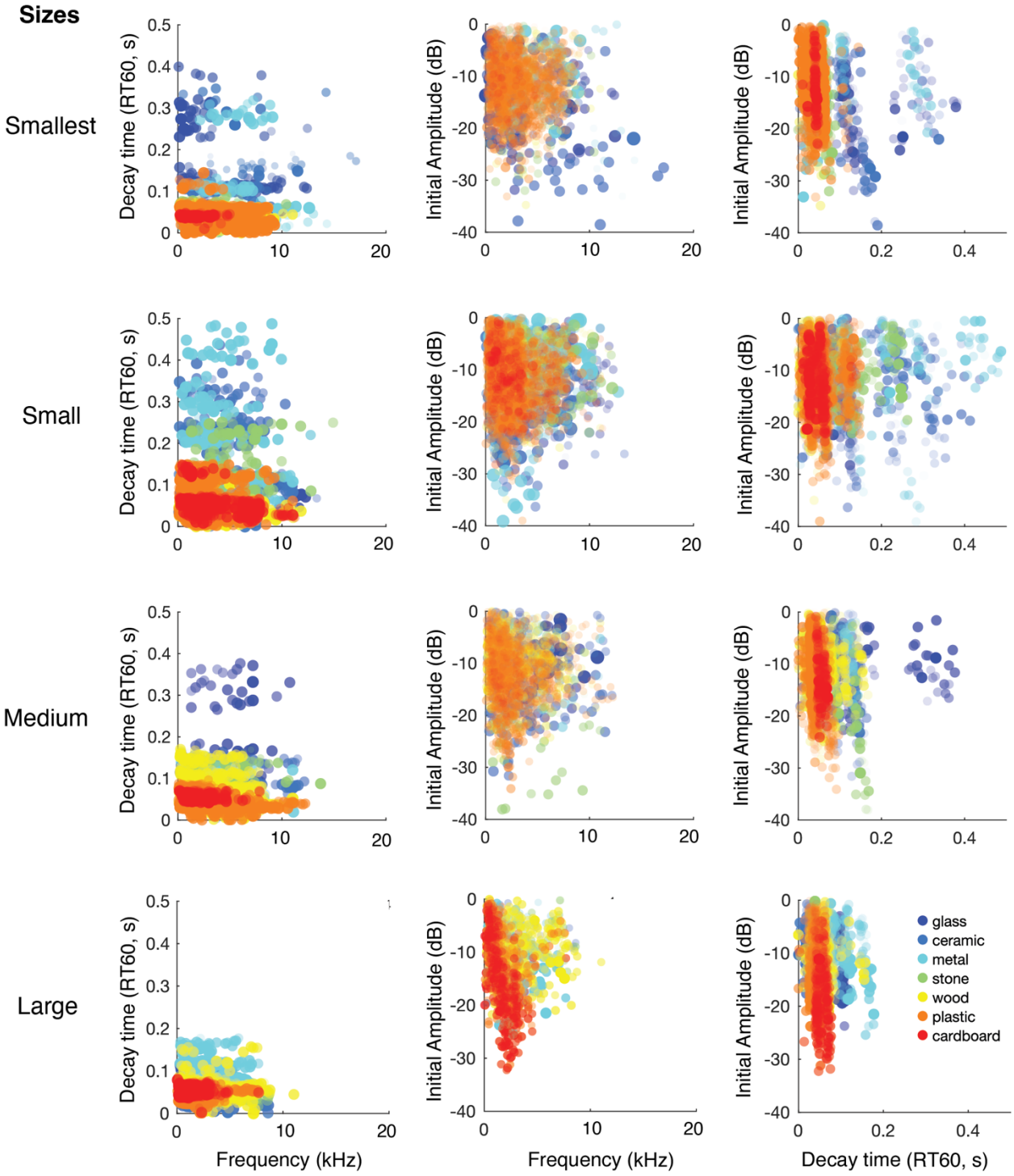



**Figure S2**. Mode parameters for all material and size categories. Each row presents the objects in one size category. Saturation of dots represents the value of the unplotted mode parameter (for instance, on the panels plotting decay time vs. frequency, the dot saturation represents initial amplitude).

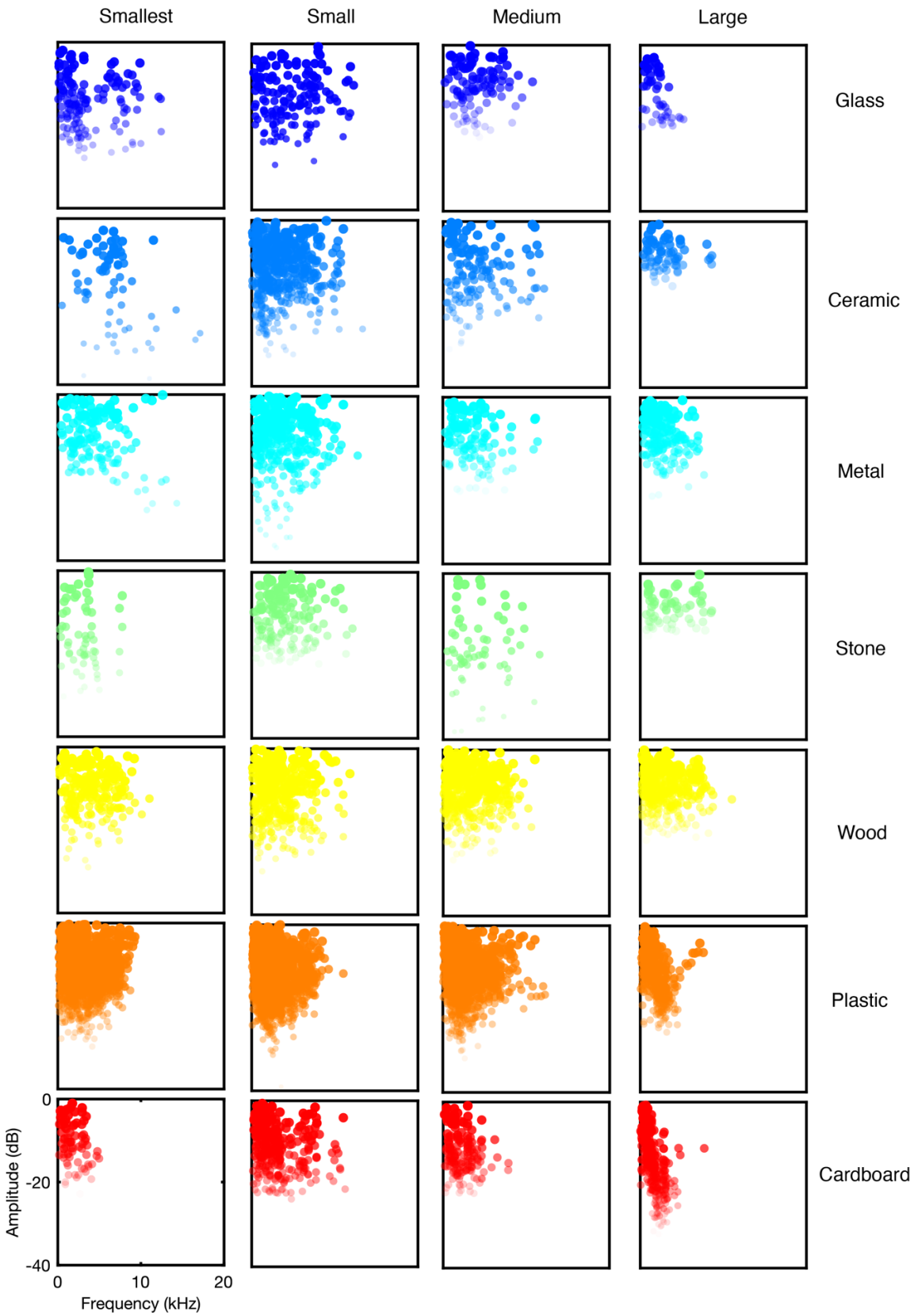



**Figure S3**. Mode amplitude and frequency for each material and size category. Same conventions as Figure S2.

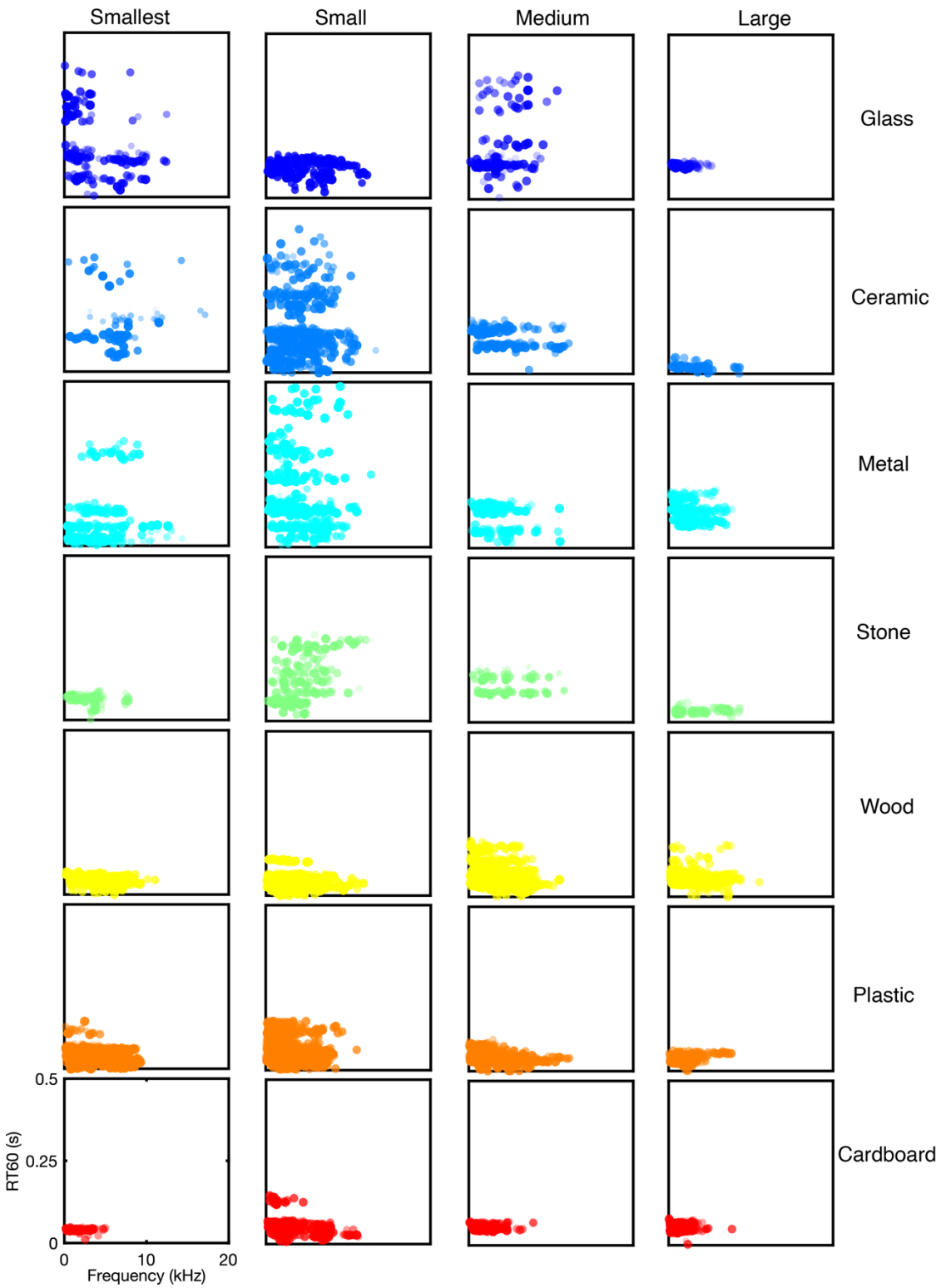



**Figure S4**. Mode decay time and frequency for each material and size category. Same conventions as Figure S2.

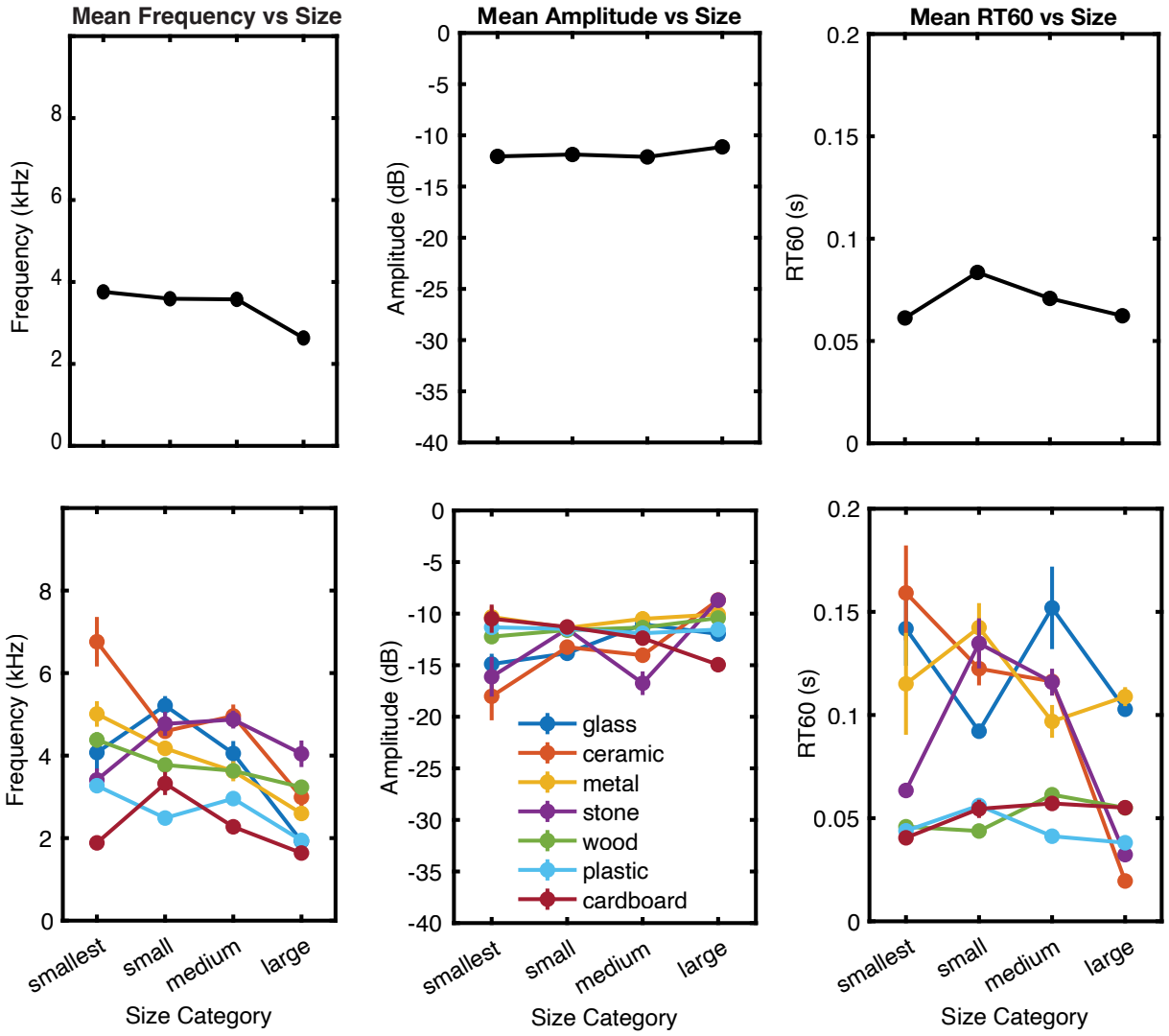



**Figure S5**. Effect of size on average mode parameters. Top shows averages across materials. Bottom shows results for individual material categories. We note that the effect of size on mode properties is in many cases confounded with object shape and damping, because different shapes and damping conditions are most common in different sizes (in the objects that humans report hearing on a daily basis). For instance, small ceramic objects are likely to be mugs, whereas large ceramic objects are likely to be tiles.

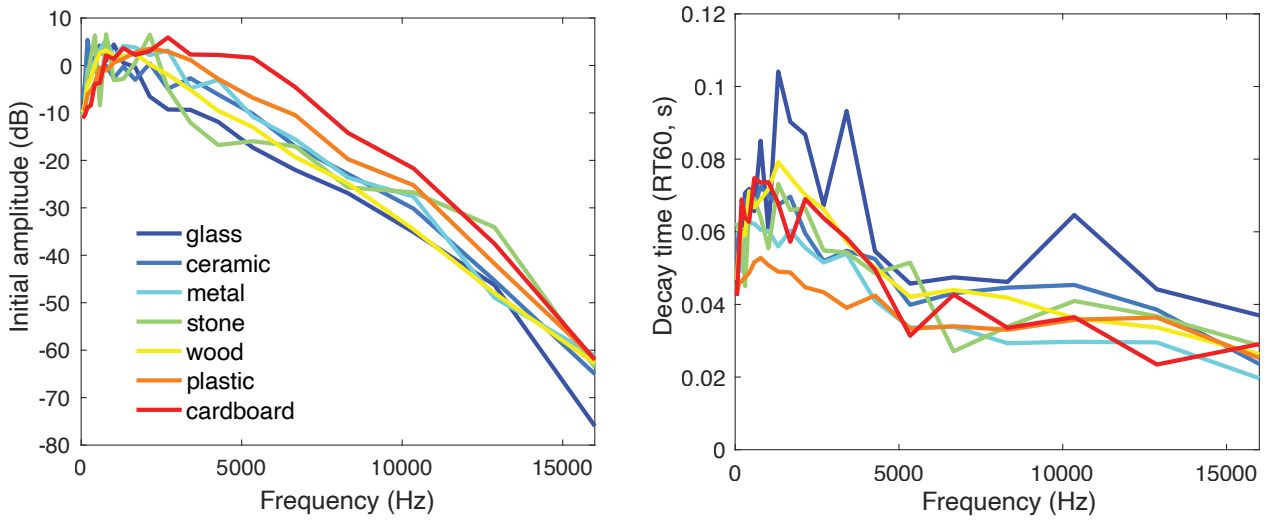



**Figure S6**. Generative parameters for noise bands of impulse responses. Note that the initial amplitudes of noise bands are higher for the two softest materials (plastic and cardboard), as expected given that these materials have less pronounced resonances.

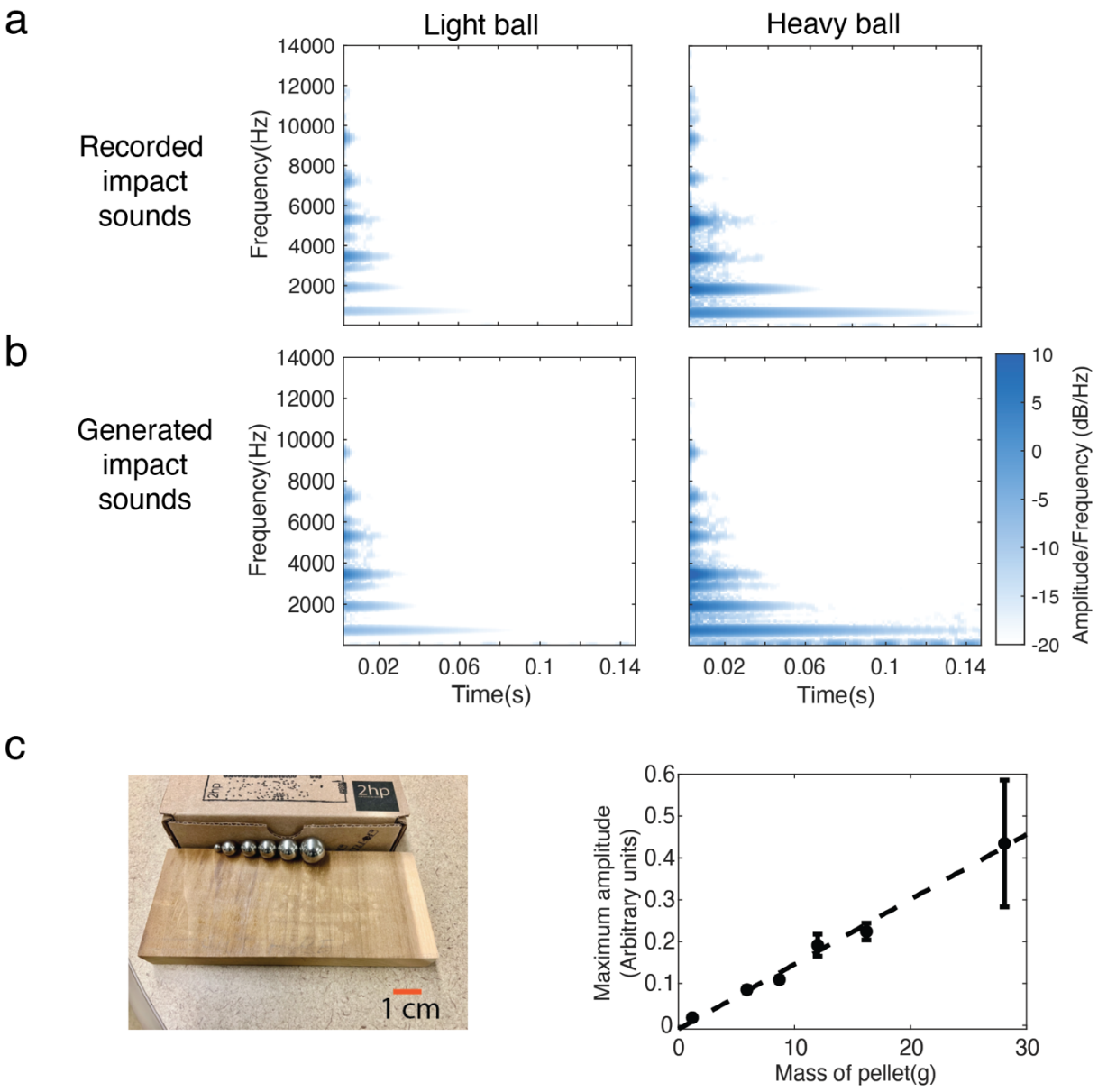



**Figure S7**. Effect of ball mass on impact sounds. **a**. Spectrograms of recorded impact sounds from different impacting object masses. The heavier mass causes the lower frequency modes of the surface to be accentuated. **b**. Spectrograms of synthesized impact sounds, generated from a simulated contact force for two different masses, using the impulse response measured from the object shown in b. The generative model successfully captures the effect of mass on the impact sound. **c**. The maximum amplitude of recorded impact sounds produced by metal pellets of different masses. Small masses dropped from modest heights produce forces that remain in the linear regime, and that increase approximately linearly with mass.

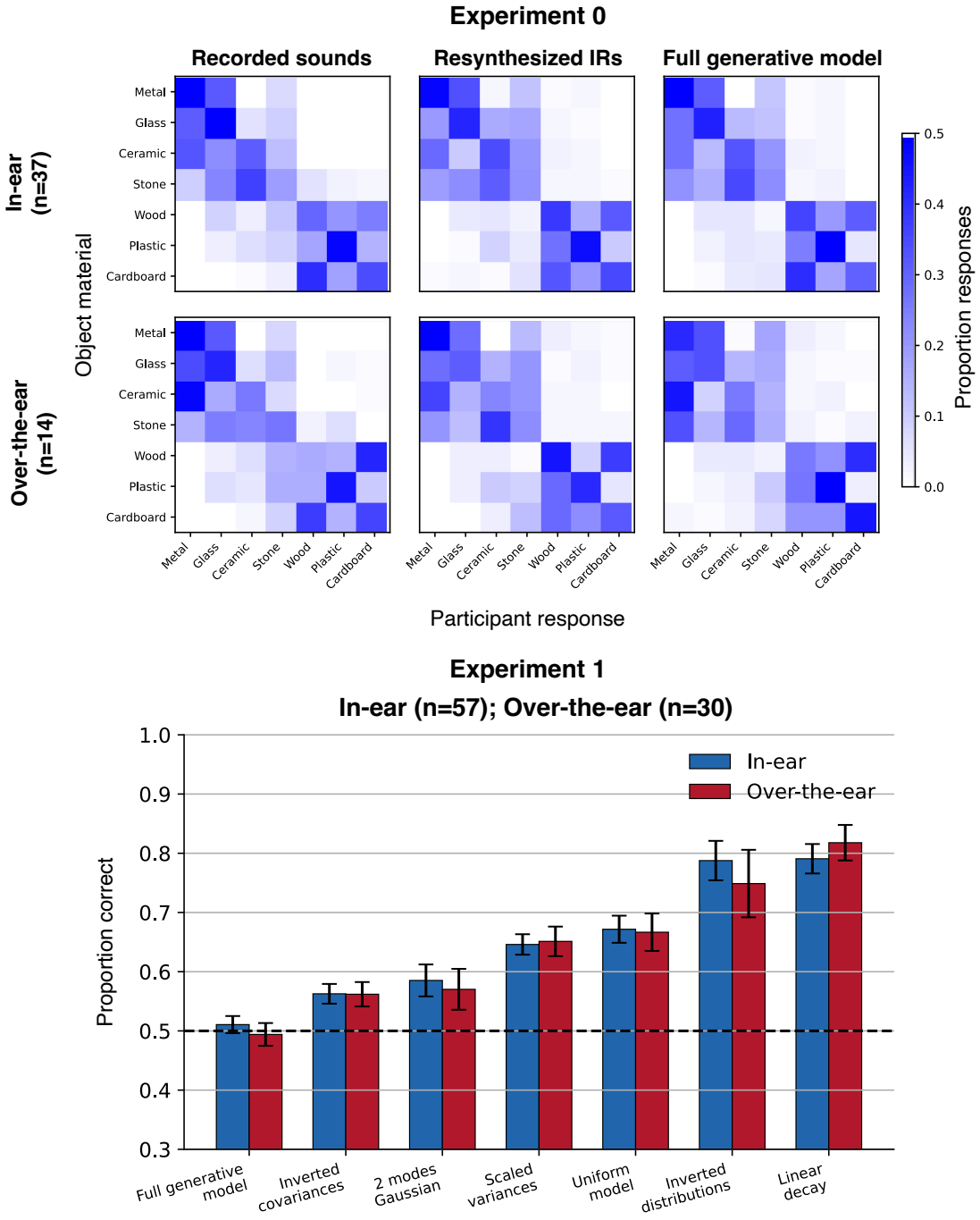



**Figure S8**. Results of Experiments 0 and 1, separated by sound presentation. Online participants self-categorized their sound presentation device as in-ear (earphones) or over-the-ear (headphones). Graphs show same data as Figures 5c and 7b, split by sound presentation type.

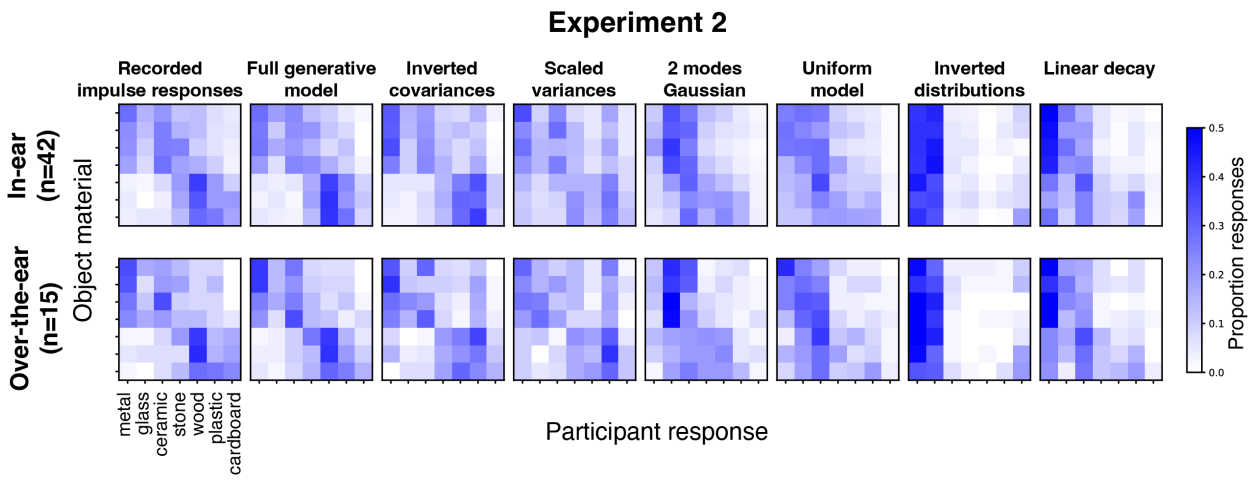



**Figure S9**. Results of Experiment 2, separated by sound presentation type. Online participants self-categorized their sound presentation device as in-ear (earphones) or over-the-ear (headphones). Graphs show same data as Figure 7d, split by sound presentation type.

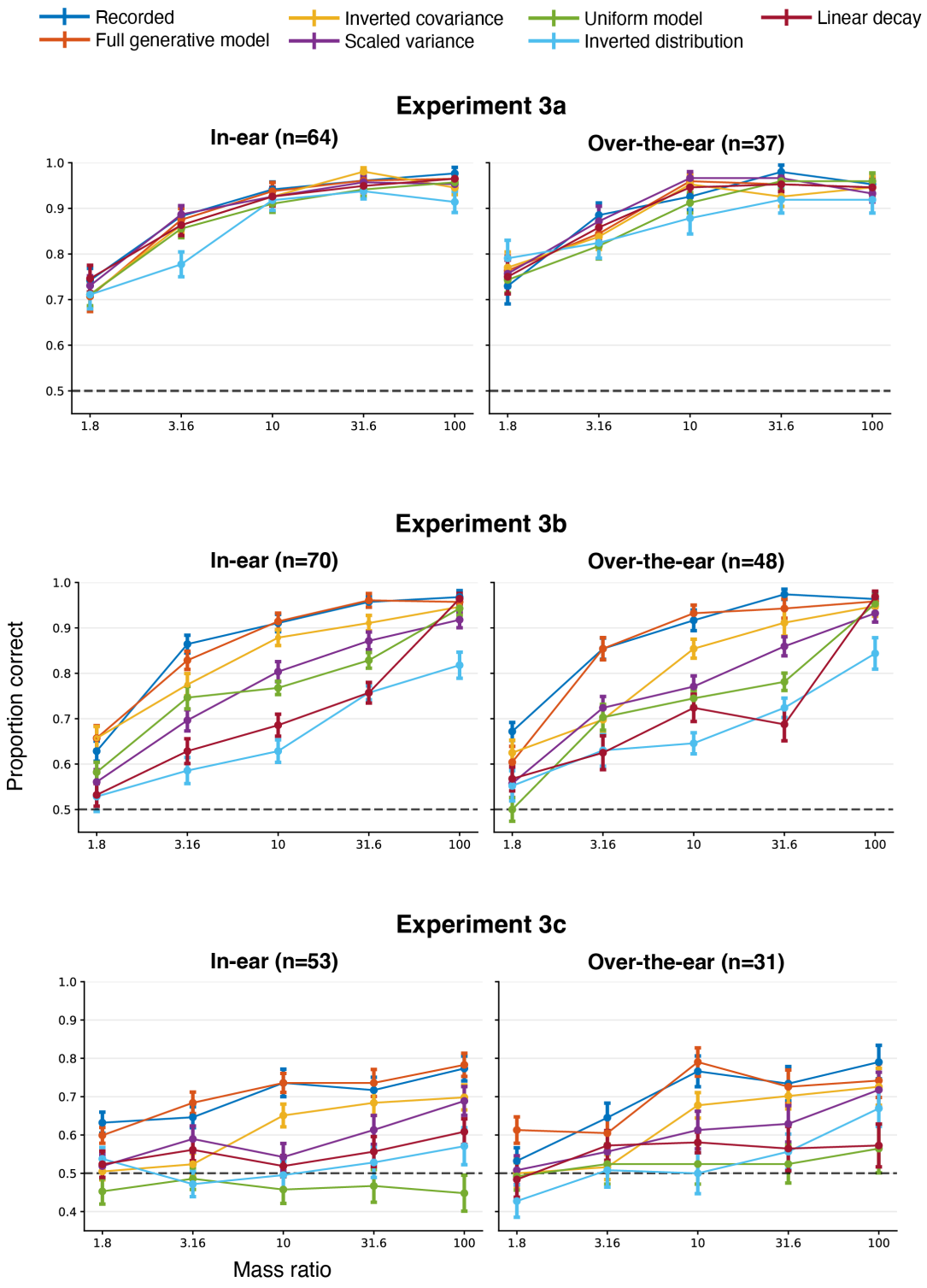



**Figure S10**. Results of Experiment 3, separated by sound presentation type. Online participants self-categorized their sound presentation device as in-ear (earphones) or over-the-ear (headphones). Top and middle graphs show same data as Figures 8g and 8h, respectively, split by sound presentation type. Bottom graph shows the data from Figure 9d along with an additional 24 participants who we recruited to boost the sample size with over-the ear headphones.

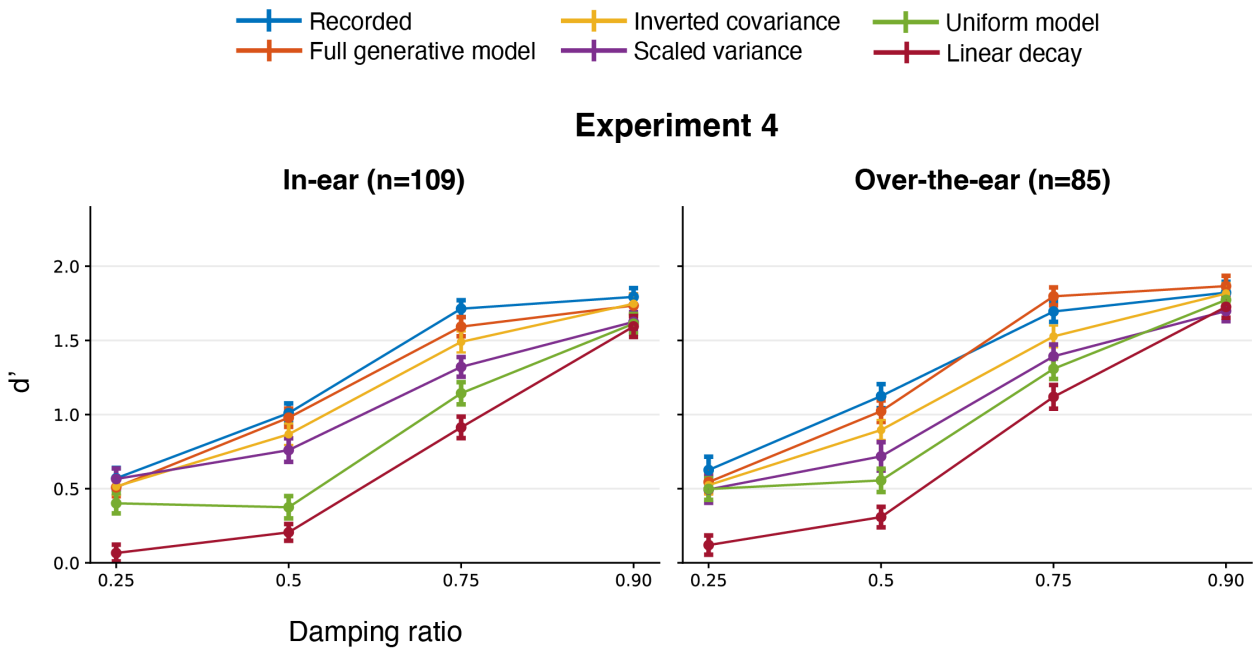



**Figure S11**. Results of Experiment 4, separated by sound presentation type. Online participants self-categorized their sound presentation device as in-ear (earphones) or over-the-ear (headphones). Graph shows the data from Figure 10e along with an additional 139 participants who we recruited to boost the sample sizes.

| Expmt. | Condition | Reference | N | P value | TestStat | Effect Size Metric | Effect Size | CI-Lower | CI-Upper |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Full generative model | Chance (0.5) | 87 | 0.67037532 | 0.43 | Cohen d: (Condition - 0.5) / SD(Condition) | 0.05 | -0.16 | 0.26 |
| 1 | Inverted covariances | Chance (0.5) | 87 | 6.21E-06 | 4.82 | Cohen d: (Condition - 0.5) / SD(Condition) | 0.52 | 0.29 | 0.78 |
| 1 | 2 modes Gaussian | Chance (0.5) | 87 | 0.00029675 | 3.77 | Cohen d: (Condition - 0.5) / SD(Condition) | 0.40 | 0.19 | 0.65 |
| 1 | Scaled variances | Chance (0.5) | 87 | 5.51E-17 | 10.45 | Cohen d: (Condition - 0.5) / SD(Condition) | 1.12 | 0.87 | 1.44 |
| 1 | Uniform model | Chance (0.5) | 87 | 1.77E-14 | 9.22 | Cohen d: (Condition - 0.5) / SD(Condition) | 0.99 | 0.75 | 1.29 |
| 1 | Inverted distributions | Chance (0.5) | 87 | 8.49E-15 | 9.37 | Cohen d: (Condition - 0.5) / SD(Condition) | 1.01 | 0.71 | 1.45 |
| 1 | Linear decay | Chance (0.5) | 87 | 8.64E-27 | 15.59 | Cohen d: (Condition - 0.5) / SD(Condition) | 1.67 | 1.33 | 2.15 |
| 2 | Inverted covariances | Full generati | 57 | 1.00E-04 | NA | r(Recorded,Full) - r(Recorded,Condition) | 0.12 | 0.04 | 0.20 |
| 2 | Uniform distribution | Full generati | 57 | 1.00E-04 | NA | r(Recorded,Full) - r(Recorded,Condition) | 0.34 | 0.22 | 0.47 |
| 2 | Two modes Gaussian | Full generati | 57 | 1.00E-04 | NA | r(Recorded,Full) - r(Recorded,Condition) | 0.70 | 0.56 | 0.79 |
| 2 | Two modes uniform | Full generati | 57 | 1.00E-04 | NA | r(Recorded,Full) - r(Recorded,Condition) | 0.66 | 0.49 | 0.75 |
| 2 | Linear decay Gaussian | Full generati | 57 | 1.00E-04 | NA | r(Recorded,Full) - r(Recorded,Condition) | 1.10 | 0.93 | 1.18 |
| 2 | Linear decay uniform | Full generati | 57 | 1.00E-04 | NA | r(Recorded,Full) - r(Recorded,Condition) | 0.72 | 0.54 | 0.83 |
| 3a | Full generative model | Recorded | 101 | 0.41585841 | NA | Threshold ratio: Condition / Recorded | 1.06 | 0.91 | 1.23 |
| 3a | Inverted covariances | Recorded | 101 | 0.28287171 | NA | Threshold ratio: Condition / Recorded | 1.07 | 0.94 | 1.24 |
| 3a | Scaled variances | Recorded | 101 | 0.99070093 | NA | Threshold ratio: Condition / Recorded | 1.00 | 0.87 | 1.18 |
| 3a | Uniform model | Recorded | 101 | 0.10368963 | NA | Threshold ratio: Condition / Recorded | 1.13 | 0.98 | 1.35 |
| 3a | Inverted distribution | Recorded | 101 | 0.0001 | NA | Threshold ratio: Condition / Recorded | 1.47 | 1.02 | 2.38 |
| 3a | Linear decay | Recorded | 101 | 0.82641736 | NA | Threshold ratio: Condition / Recorded | 1.02 | 0.84 | 1.20 |
| 3b | Full generative model | Recorded | 118 | 0.26717328 | NA | Threshold ratio: Condition / Recorded | 1.04 | 0.97 | 1.13 |
| 3b | Inverted covariances | Recorded | 118 | 1.00E-04 | NA | Threshold ratio: Condition / Recorded | 2.25 | 1.63 | 2.74 |
| 3b | Scaled variances | Recorded | 118 | 1.00E-04 | NA | Threshold ratio: Condition / Recorded | 4.59 | 3.07 | 6.99 |
| 3b | Uniform model | Recorded | 118 | 1.00E-04 | NA | Threshold ratio: Condition / Recorded | 9.95 | 7.08 | 13.94 |
| 3b | Inverted distribution | Recorded | 118 | 1.00E-04 | NA | Threshold ratio: Condition / Recorded | 27.75 | 20.56 | Inf |
| 3b | Linear decay | Recorded | 118 | 1.00E-04 | NA | Threshold ratio: Condition / Recorded | 18.76 | 15.43 | 21.49 |
| 3c | Full generative model | Recorded | 60 | 0.25077492 | NA | Mean performance: Condition - Recorded | 0.02 | -0.01 | 0.05 |
| 3c | Inverted covariances | Recorded | 60 | 1.00E-04 | NA | Mean performance: Condition - Recorded | -0.07 | -0.10 | -0.04 |
| 3c | Scaled variances | Recorded | 60 | 1.00E-04 | NA | Mean performance: Condition - Recorded | -0.12 | -0.15 | -0.08 |
| 3c | Uniform model | Recorded | 60 | 1.00E-04 | NA | Mean performance: Condition - Recorded | -0.22 | -0.26 | -0.19 |
| 3c | Inverted distribution | Recorded | 60 | 1.00E-04 | NA | Mean performance: Condition - Recorded | -0.19 | -0.23 | -0.16 |
| 3c | Linear decay | Recorded | 60 | 1.00E-04 | NA | Mean performance: Condition - Recorded | -0.18 | -0.21 | -0.15 |
| 4 | Full generative model | Recorded | 55 | 0.41485851 | NA | Threshold ratio: Condition / Recorded | 1.05 | 0.94 | 1.19 |
| 4 | Inverted covariances | Recorded | 55 | 0.00389961 | NA | Threshold ratio: Condition / Recorded | 1.23 | 1.07 | 1.47 |
| 4 | Scaled variances | Recorded | 55 | 1.00E-04 | NA | Threshold ratio: Condition / Recorded | 1.31 | 1.18 | 1.48 |
| 4 | Uniform distribution | Recorded | 55 | 1.00E-04 | NA | Threshold ratio: Condition / Recorded | 1.35 | 1.22 | 1.53 |
| 4 | Linear decay | Recorded | 55 | 0.00019998 | NA | Threshold ratio: Condition / Recorded | 1.40 | 1.22 | 1.66 |

**Table S5**. Full statistical reporting for Experiments 1-4. P values for Experiments 2-4 are obtained through permutation tests, and so have a minimum value determined by the number of permutations.

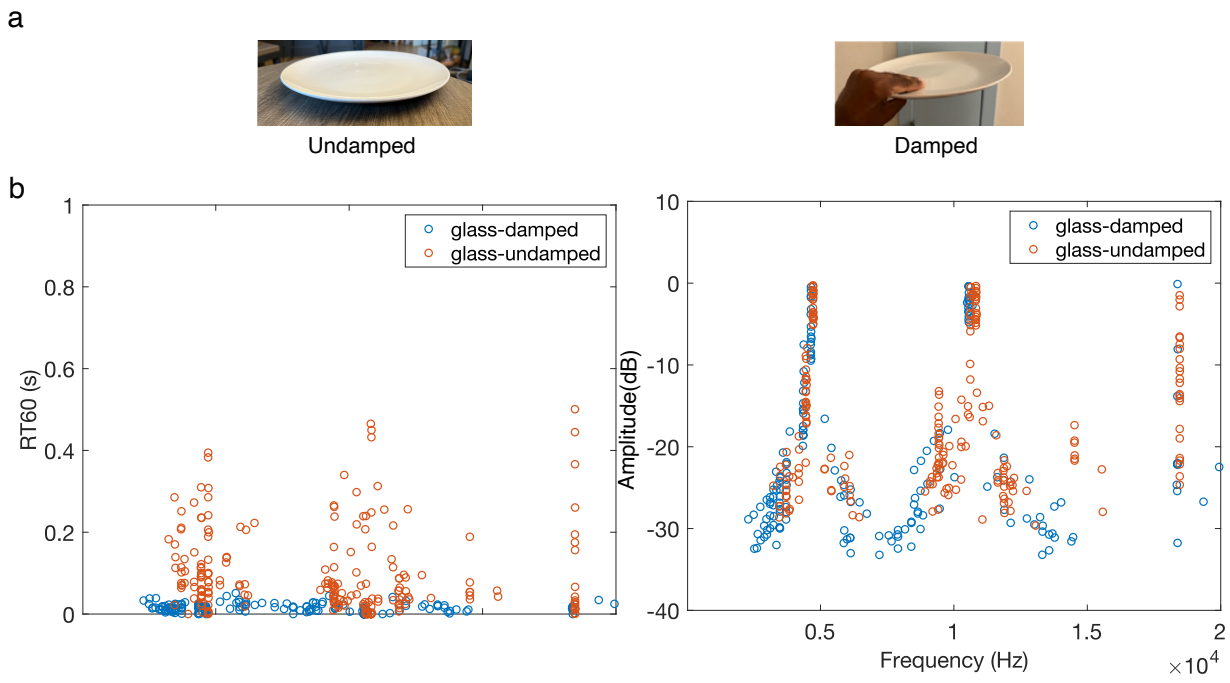



**Figure S12**. Acoustic effects of everyday object damping. We recorded impacts produced by tapping a plate that either rested freely on a table or was held by a human hand. Graphs plot the decay time, amplitude and frequency of modes measured from the recorded impacts. Decay times are much shorter for the damped plates. Frequencies and amplitudes, by contrast, remain similar.

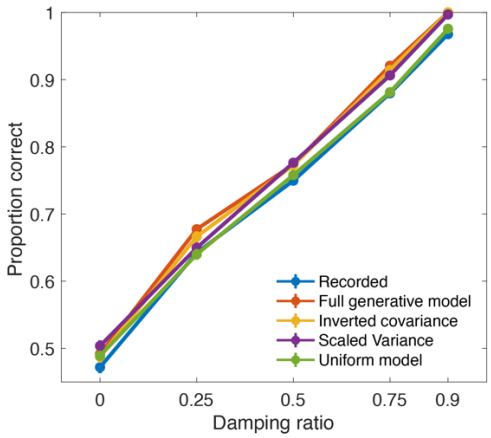



**Figure S13**. Observer model performance for damping discrimination (Experiment 4). Graph plots performance of observer model that selected the stimulus with lower decay time (average RT60 across modes) as being damped. Error bars plot SEM across 200 simulated experiment runs, and are smaller than the dot symbols. The purpose of the model was to verify that the experimental conditions did not differ in the cue that might be supposed to support performance on the experimental task. It is not intended as a model of human perception.

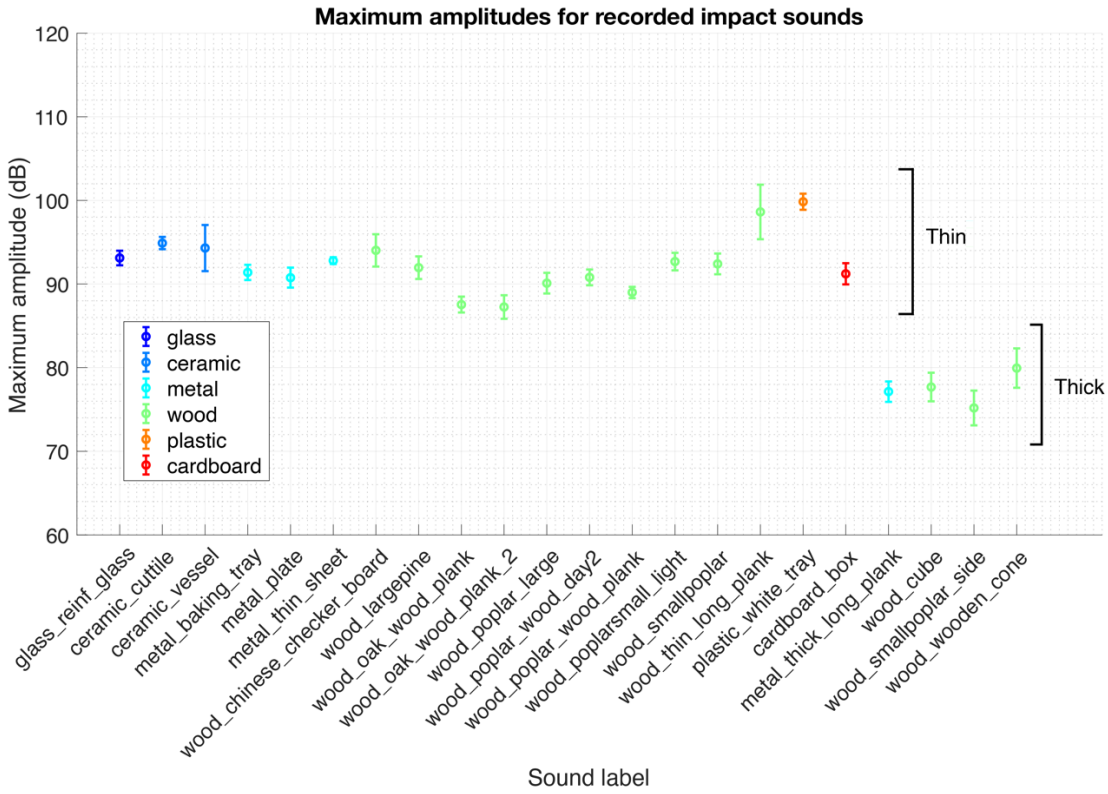



**Figure S14**. Mode amplitudes are similar across different materials. Because the impulse response measurements did not preserve absolute levels, we conducted an experiment to assess whether level varied across materials. We recorded impact sounds from objects and measured the initial amplitude of each sound. Because the sound amplitude is also influenced by the amplitude of the force of impact, we held this constant by fixing the incident momentum (by using a fixed pellet dropped from a fixed height above the resonating surface). Sounds were recorded with an external microphone, with the gain on the sound interface held constant across recordings. We estimated the initial sound amplitude to be the maximum value of the measured sound waveform. The graph plots the mean amplitude for each object computed over the 10-15 impacts we recorded per object. Error bars plot SD. Amplitudes did not systematically vary with material. The main factor we found to influence amplitude was surface thickness, with thicker surfaces having lower initial amplitudes than thinner surfaces. Given that our analysis and experiments focused on effects of material (and overall object size), this experiment suggests that the impulse response normalization did not discard information that was critical for our purposes.

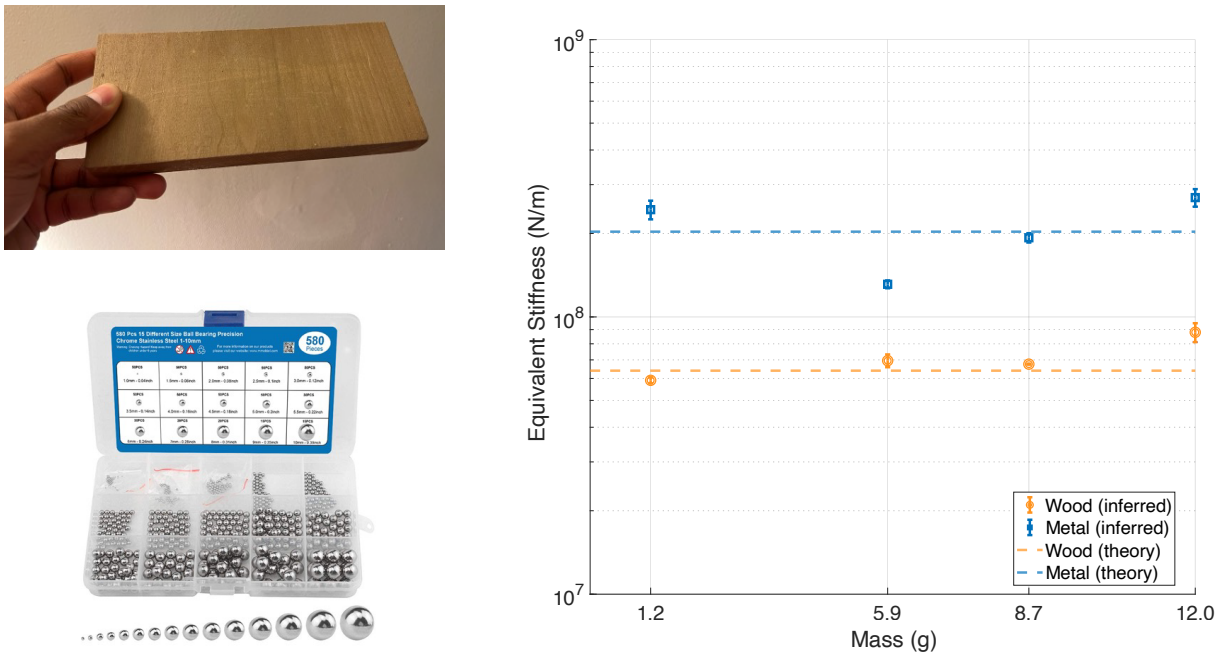



**Figure S15**. Validation of stiffness values and forcing function assumptions. Dots plot inferred stiffness for different impact masses and surface materials. Stiffness was inferred by fitting the generative model to the audio waveform recorded from impacts of small metal pellets on wood and metal surfaces. Dashed lines plot stiffness values used in the generative model for wood and metal in the small size category (given in Table 3 based on theoretical derivation using the constants in Table 2). See Methods for details.

## Supplementary References


1 Foster, S. in *IEEE International Conference on Acoustics, Speech, and Signal Processing* Vol. 11 929-932 (IEEE, 1986).

2 Hess, R. L. *Transverse Vibration of Beams* PhD thesis, University of Michigan, (1950).

3 Gaver, W. W. How do we hear in the world?: Explorations in ecological acoustics. *Ecological Psychology* **5**, 285-313 (1993).

4 Aramaki, M., Kronland-Martinet, R., Voinier, T. & Ystad, S. A percussive sound synthesizer based on physical and perceptual attributes. *Computer Music Journal* **30**, 32-41 (2006).

5 Ren, Z., Yeh, H. & Lin, M. C. Example-guided physically based modal sound synthesis. *ACM Transactions on Graphics (TOG)* **32** (2013).

6 Traer, J., Cusimano, M. & McDermott, J. H. in *The 22nd International Conference on Digital Audio Effects (DAFx-19)* (2019).

7 Glasberg, B. R. & Moore, B. C. J. Derivation of auditory filter shapes from notched-noise data. *Hearing Research* **47**, 103-138, doi:10.1016/0378-5955(90)90170-T (1990).

8 Huber, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35**, 73-101 (1964).

9 Rayleigh, J. W. S. B. *The theory of sound*. (Macmillan, 1896).

10 Johnston, J. D. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications* **6**, 314-323, doi:10.1109/49.608 (1988).

11 Grassi, M. Do we hear size or sound? Balls dropped on plates. *Perception & Psychophysics* **67**, 274-284 (2005).

12 Grassi, M., Pastore, M. & Lemaitre, G. Looking at the world with your ears: How do we get the size of an object from its sound? *Acta Psychologica* **143**, 96-104 (2013).

13 Johnson, K. L., Kendall, K. & Roberts, A. D. Surface energy and the contact of elastic solids. *Proceedings of the Royal Society A* **324**, 301-313 (1971).

14 Johnson, K. L. *Contact Mechanics*. (Cambridge University Press, 1985).

15 Reid, S. R. & Peng, C. Dynamic uniaxial crushing of wood. *International Journal of Impact Engineering* **19**, 531-570 (1997).

16 Warren, W. H. & Verbrugge, R. R. Auditory perception of breaking and bouncing events: A case study in ecological acoustics. *Journal of Experimental Psychology: Human Perception and Performance* **10**, 704-712 (1984).

17 Traer, J., Norman-Haignere, S. V. & McDermott, J. H. Causal inference in environmental sound recognition. *Cognition* **214**, 104627 (2021).

18 Woods, K. J. P. & McDermott, J. Schema learning for the cocktail party problem. *Proceedings of the National Academy of Sciences* **115**, E3313-E3322 (2018).

19 Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. & McDermott, J. H. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630-644, doi:10.1016/j.neuron.2018.03.044 (2018).

20 McWalter, R. & McDermott, J. H. Illusory sound texture reveals multi-second statistical completion in auditory scene analysis. *Nature Communications* **10**, 5096 (2019).

21 McPherson, M. J. & McDermott, J. H. Time-dependent discrimination advantages for harmonic sounds suggest efficient coding for memory. *Proceedings of the National Academy of Sciences* **117**, 32169-32180, doi:10.1101/2020.05.07.082511 (2020).

22 McPherson, M. J., Dolan, S. E., Durango, A., Ossandon, T., Valdés, J., Undurraga, E. A., Jacoby, N., Godoy, R. A. & McDermott, J. H. Perceptual fusion of musical notes by native

Amazonians suggests universal representations of musical intervals. *Nature Communications* **11**, 2786 (2020).
23 McPherson, M. J., Grace, R. C. & McDermott, J. H. Harmonicity aids hearing in noise. *Attention, Perception, and Psychophysics* **84**, 1016-1042 (2022).
24 Woods, K. J. P., Siegel, M. H., Traer, J. & McDermott, J. H. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics* **79**, 2064-2072 (2017).