A Modal Synthesis Background

We adopt tetrahedral finite element models to represent any given geometry [O'Brien et al. 2002]. The displacements, $\mathbf{x} \in \mathbb{R}^{3N}$, in such a system can be calculated with the following linear deformation equation:

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{C}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f}, \tag{1}$$

where $\mathbf{M}$, $\mathbf{C}$, and $\mathbf{K}$ respectively represent the mass, damping and stiffness matrices. We approximate the damping matrix with *Rayleigh damping*: $\mathbf{C} = \alpha\mathbf{M} + \beta\mathbf{K}$, which is a well-established practice. The system can be decoupled into the following form:

$$\ddot{\mathbf{q}} + (\alpha\mathbf{I} + \beta\mathbf{\Lambda})\dot{\mathbf{q}} + \mathbf{\Lambda}\mathbf{q} = \mathbf{U}^T\mathbf{f}, \tag{2}$$

where $\mathbf{\Lambda}$ is a diagonal matrix.The solution to Eqn. 2 is a bank of *modes*, i.e. damped sinusoidal waves. The i'th mode is

$$q_i = a_i e^{-d_i t}\sin(2\pi f_i t + \theta_i), \tag{3}$$

where f_i is the frequency of the mode, d_i is the damping coefficient, a_i is the excited amplitude, and θ_i is the initial phase. (f_i, d_i, a_i) together define the *feature* of mode i.

The values in Eqn. 3 depend on the material properties, the geometry, and the run-time interactions: a_i and θ_i depend on the run-time excitation of the object, while f_i and d_i depend on the geometry and the material properties:

$$d_i = \frac{1}{2}(\alpha + \beta\lambda_i), \tag{4}$$

$$f_i = \frac{1}{2\pi}\sqrt{\lambda_i - \left(\frac{\alpha + \beta\lambda_i}{2}\right)^2}. \tag{5}$$

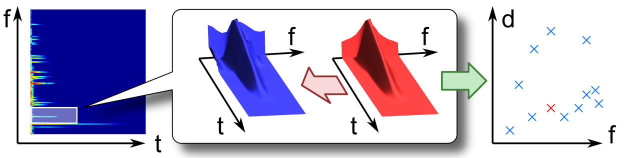
where the eigenvalues λ_i's are calculated from $\mathbf{M}$ and $\mathbf{K}$, which in turn depend on mass density ρ, Young's modulus E, and Poisson's ratio ν.

B Feature Extraction

We extract the features $\{f_i, d_i, a_i\}$ from the example audio using a time-varying frequency representation called *power spectrogram*. A power spectrogram $\mathbf{P}$ for a a time domain signal $\mathbf{s}[n]$, is obtained by first breaking it up into overlapping frames, and then performing windowing and Fourier transform on each frame:

$$\mathbf{P}[m, \omega] = \left|\sum_n \mathbf{s}[n]\mathbf{w}[n-m]e^{-j\omega n}\right|^2, \tag{6}$$

where $\mathbf{w}$ is the window applied to the original time domain signal.



(a) *select range* **(b)** *fit local shape* **(c)** *extracted features*

Figure 2

The features are then extracted from the power spectrogram through the process shown in Figure 2. First, a peak is detected in a power spectrogram at the location of a potential mode (Figure 2a, where f=frequency, t=time). Then a local shape fitting of the power spectrogram is performed to estimate the frequency, damping and amplitude of the potential mode (Figure 2b). Finally, if the fitting error is below a certain threshold, we collect it in the set of extracted features, shown as the red cross in the feature space (Figure 2c, where only the frequency f and damping d are shown).

C Parameter Estimation

C.1 Optimization Framework

The material parameters are estimated through an optimization framework. We first create a virtual object that is roughly the same size and geometry as the real-world object whose impact sound was recorded. We then calculate its mass matrix $\mathbf{M}$ and stiffness matrix $\mathbf{K}$ and find the assumed eigenvalues λ_i^0's using some initial values for the Young's modulus, mass density, and Poisson's ratio, E_0, ρ_0, and ν_0. The eigenvalue λ_i for general E and ρ is just a multiple of λ_i^0:

$$\lambda_i = \frac{\gamma}{\gamma_0}\lambda_i^0 \tag{7}$$

where $\gamma = E/\rho$ is the ratio of Young's modulus to density, and $\gamma_0 = E_0/\rho_0$ is the ratio using the assumed values. Applying a unit impulse on the virtual object at a point corresponding to the actual impact point in the example recording gives an excitation pattern of the eigenvalues as Eqn. 3, where the excitation amplitude of mode j is a_j^0. If the actual (unknown) impulse is not unit, then the excitation amplitude is just scaled by a factor σ,

$$a_j = \sigma a_j^0 \tag{8}$$

2 Combining Eqn. 4, Eqn. 5, Eqn.7, and Eqn.8, we obtain a mapping from an assumed eigenvalue and its excitation (λ_j^0, a_j^0) to an estimated mode with frequency $\tilde{f}_j$, damping $\tilde{d}_j$, and amplitude $\tilde{a}_j$:

$$(\lambda_j^0, a_j^0) \xrightarrow{\{\alpha,\beta,\gamma,\sigma\}} (\tilde{f}_j, \tilde{d}_j, \tilde{a}_j). \tag{9}$$

The estimated sound $\tilde{\mathbf{s}}[n]$ is generated by mixing all the estimated modes,

$$\tilde{\mathbf{s}}[n] = \sum_j \left(\tilde{a}_j e^{-\tilde{d}_j(n/F_s)}\sin(2\pi\tilde{f}_j(n/F_s))\right) \tag{10}$$

where F_s is the sampling rate.

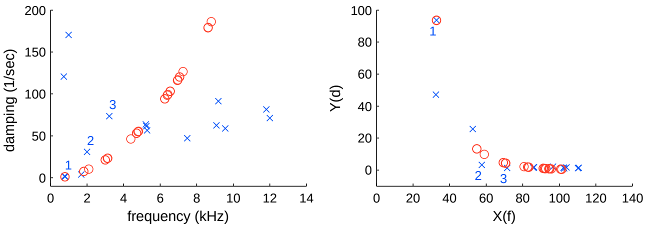
The estimated sound $\tilde{\mathbf{s}}[n]$ can then be compared against the example sound $\mathbf{s}[n]$ and a difference metric can be computed. An optimization process is used to find the parameter set with minimal difference metric value.

C.2 Psychoacoustic Metric

A combination of two metrics is used: an 'image domain metric' that evaluates the perceptual similarity of sound clips, and a 'feature domain metric' that measures the audio material resemblance.

Image Domain Metric: Given an reference sound $\mathbf{s}[n]$ and an estimated sound $\tilde{\mathbf{s}}[n]$, their power spectrograms are computed using Eqn. 6. The power spectrograms are transformed before the difference is taken. The frequency axis is transformed to *critical band rate* z to account for humans' better ability to distinguish lower frequencies than higher frequencies [Zwicker and Fastl 1999]. The intensity is transformed from pressure to *loudness*, a perceptual value that measures human sensation to sound intensity.

Feature Domain Metric: To measure the resemblance between extracted (real) features and estimated (synthesized) features, we use a point set matching metric. First the frequency and damping of feature points, (f, d), are transformed. The frequency is transformed to the critical band rate as described previously. The damping is transformed to duration, which is proportional to the inverse of the damping value. Figure 3 shows the effect of the transformation. A matching score can then be computed between the transformed point sets.



(a) **(b)**

Figure 3: *Point set matching problem in the feature domain: (a) in the original frequency and damping, (f, d)-space. (b) in the transformed, (x, y)-space, where $x = X(f)$ and $y = Y(d)$. The blue crosses and red circles are the reference and estimated feature points respectively. The three features having the largest energies are labeled 1, 2, and 3.*

D Residual Compensation

D.1 Residual Computation

Figure 4 illustrates the residual computation process. From a recorded sound (Figure 4a), the reference features are extracted (Figure 4b), with frequencies, dampings, and energies depicted as the blue circles (Figure 4f). After parameter estimation, the synthesized sound is generated (Figure 4c), with the estimated features shown as the red crosses (Figure 4g), which all lie on a curve in the (f, d)-plane. Each reference feature may be approximated by one or more estimated features, and its match ratio number is shown. The represented sound is the summation of the reference features weighted by their match scores, shown as the solid blue circles (Figure 4h). Finally, the difference between the recorded sound's power spectrogram (Figure 4a) and the represented sound's (Figure 4d) are computed to obtain the residual (Figure 4e).

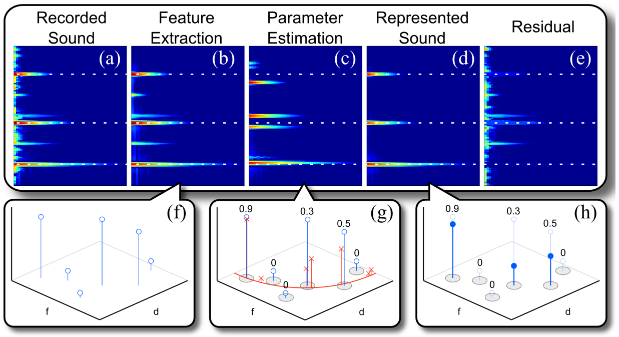


Figure 4: *Residual computation.*

D.2 Residual Transfer

As discussed in previous sections, *modes* transfer naturally with geometries in the modal analysis process, and they respond to excitations at runtime in a physical manner. In other words, the modal component of the synthesized sounds already provides transferability of sounds due to varying geometries and dynamics. Hence, we compute the transferred residual under the guidance of modes. Algorithm 1 shows the complete feature-guided residual transfer algorithm.

Algorithm 1: Residual Transformation at Runtime

```
Input: source modes $\mathbf{\Phi}^s = \{\phi_i^s\}$, target modes $\mathbf{\Phi}^t = \{\phi_j^t\}$, and
       source residual audio $\mathbf{s}^s_{residual}[n]$
Output: target residual audio $\mathbf{s}^t_{residual}[n]$
$\mathbf{\Psi} \leftarrow$ DetermineModePairs($\mathbf{\Phi}^s, \mathbf{\Phi}^t$)
foreach mode pair $(\phi_k^s, \phi_k^t) \in \mathbf{\Psi}$ do
    $\mathbf{P}^{s\prime} \leftarrow$ ShiftSpectrogram( $\mathbf{P}^s$, $\Delta$frequency)
    $\mathbf{P}^{s\prime\prime} \leftarrow$ StretchSpectrogram( $\mathbf{P}^{s\prime}$, damping_ratio)
    $\mathbf{A} \leftarrow$ FindPixelScale($\mathbf{P}^t, \mathbf{P}^{s\prime\prime}$)
    ${\mathbf{P}^s_{residual}}' \leftarrow$ ShiftSpectrogram($\mathbf{P}^s_{residual}$, $\Delta$frequency)
    ${\mathbf{P}^s_{residual}}'' \leftarrow$ StretchSpectrogram(${\mathbf{P}^s_{residual}}'$, damping_ratio)
    ${\mathbf{P}^t_{residual}}'' \leftarrow$ MultiplyPixelScale(${\mathbf{P}^s_{residual}}''$, $\mathbf{A}$)
    $(\omega_{start}, \omega_{end}) \leftarrow$ FindFrequencyRange($\phi^t_{k-1}, \phi^t_k$)
    $\mathbf{P}^t_{residual}$ [m, $\omega_{start}, \ldots, \omega_{end}$] $\leftarrow$ ${\mathbf{P}^t_{residual}}''$ [m, $\omega_{start}, \ldots, \omega_{end}$]
end
$\mathbf{s}^t_{residual}[n] \leftarrow$ IterativeInverseSTFT($\mathbf{P}^t_{residual}$)
```

E Results

Parameter estimation: We estimate the material parameters from various real-world audio recordings: a wood plate, a plastic plate, a metal plate, a porcelain plate, and a glass bowl. For each recording, the parameters are estimated using a virtual object that is of the same size and shape as the one used to record the audio clips. When the virtual object is hit at the same location as the real-world object, it produces a sound similar to the recorded audio, as shown in Fig. 5 and the supplementary video.

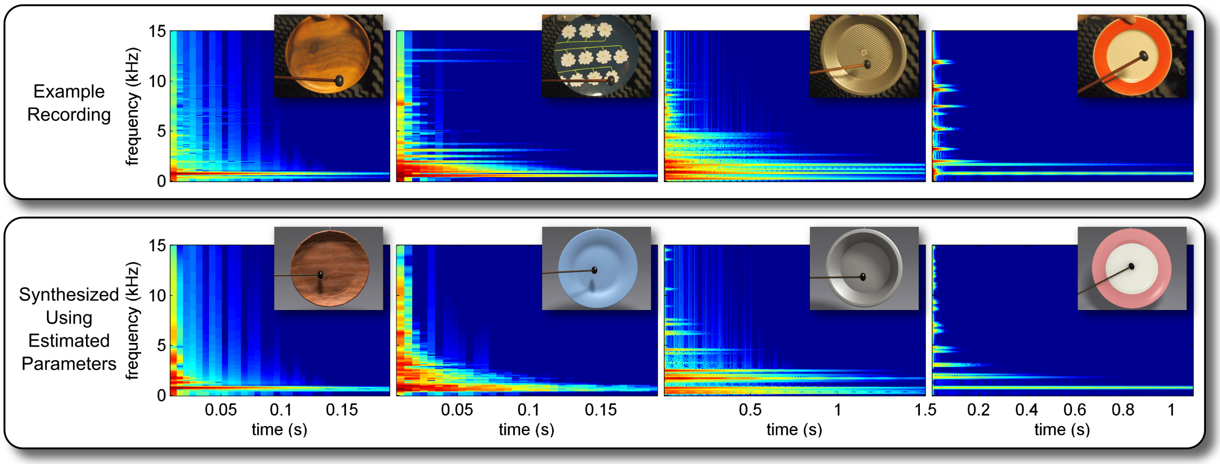
Fig. 6 compares the refenece features of the real-world objects and the estimated features of the virtual objects as a result of the parameter estimation.

Transfered parameters and residual: The parameters estimated as well as the residuals can be transfered to virtual objects with different sizes and shapes as shown in Fig. 7. From an example recording of a porcelain plate (a), the parameters for the porcelain material are estimated, and the residual computed (b). The parameters and residual are then transfered to a smaller porcelain plate (c) and a porcelain bunny (d).

Comparison with real recordings: Fig. 8 shows a comparison of the transferred results with the real recordings. From a recording of glass bowl, the parameters for glass are estimated (column (a)) and transfered to other virtual glass bowls of different sizes. The synthesized sounds ((b) (c) (d), bottom row) are compared with the real-world audio for these different-sized glass bowls ((b) (c) (d), top row). More examples of transferring the material parameters as well as the residuals are demonstrated in the supplementary video.

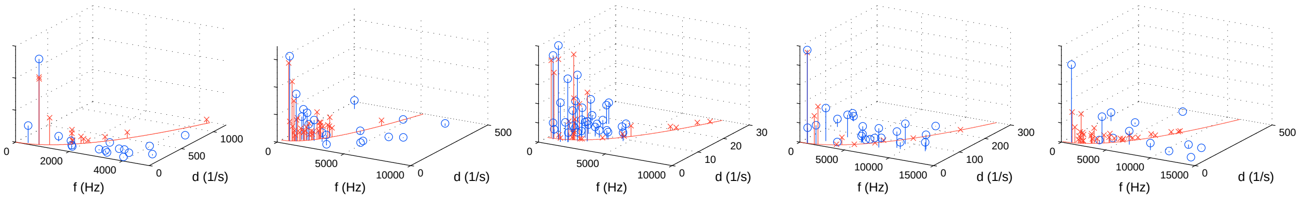
F Perceptual Study

we also designed an experiment to evaluate the auditory perception of the synthesized sounds of five different materials. Each subject is presented with a series of 24 audio clips: 8 are audio recordings of sound generated from hitting a real-world objec; 16 are synthesized using the techniques described in this paper. For each audio clip,



(a) *wood plate* (b) *plastic plate* (c) *metal plate* (d) *porcelain plate*

Figure 5: *Parameter estimation for different materials. For each material, the material parameters are estimated using an example recorded audio (top row). Applying the estimated parameters to a virtual object with the same geometry as the real object used in recording the audio will produce a similar sound (bottom row).*



(a) *wood plate* **(b)** *plastic plate* **(c)** *metal plate* **(d)** *porcelain plate* **(e)** *glass bowl*

Figure 6: *Feature comparison of real and virtual objects. The blue circles represent the reference features extracted from the recordings of the real objects. The red crosses are the features of the virtual objects using the estimated parameters. Because of the Rayleigh damping model, all the features of a virtual object lie on the depicted red curve on the* (f, d)*-plane.*

	Recognized Material				
Recorded Material	Wood (%)	Plastic (%)	Metal (%)	Porcelain (%)	Glass (%)
Wood	50.7	47.9	0.0	0.0	1.4
Plastic	37.5	37.5	6.3	0.0	18.8
Metal	0.0	0.0	66.1	9.7	24.2
Porcelain	0.0	0.0	1.2	15.1	83.7
Glass	1.7	1.7	1.7	21.6	73.3

Table 1: *Material Recognition Rate Matrix: Recorded Sounds*

	Recognized Material				
Synthesized Material	Wood (%)	Plastic (%)	Metal (%)	Porcelain (%)	Glass (%)
Wood	52.8	43.5	0.0	0.0	3.7
Plastic	43.0	52.7	0.0	2.2	2.2
Metal	1.8	1.8	69.6	15.2	11.7
Porcelain	0.0	1.1	7.4	29.8	61.7
Glass	3.3	3.3	3.8	40.4	49.2

Table 2: *Material Recognition Rate Matrix: Synthesized Sounds Using Our Method*

the subject is asked to identify among a set of 5 choices (wood, plastic, metal, porcelain, and glass), from which the sound came.

Table 1 presents the recognition rates of sounds from real-world materials, and Table 2 reflects the recognition rates of sounds from synthesized virtual materials. We found that the successful recognition rate of virtual materials using our synthesized sounds compares favorably to the recognition rate of real materials using recorded sounds.

References

O'Brien, J. F., Shen, C., and Gatchalian, C. M. 2002. Synthesizing sounds from rigid-body simulations. In *The ACM SIGGRAPH 2002 Symposium on Computer Animation*, ACM Press, 175–181.

Zwicker, E., and Fastl, H. 1999. *Psychoacoustics: Facts and models*, 2nd updated edition ed., vol. 254. Springer New York.

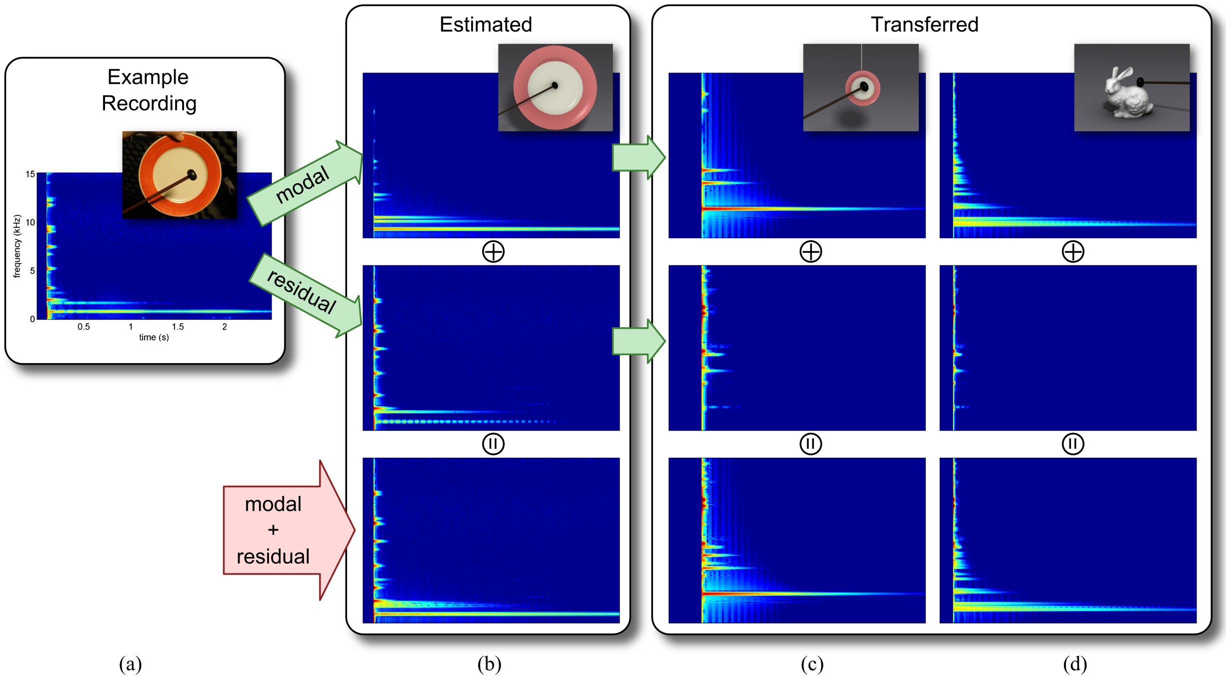


Figure 7: *Transfered material parameters and residual: from a real-world recording (a), the material parameters are estimated and the residual computed (b). The parameters and residual can then be applied to various objects made of the same material, including (c) a smaller object with similar shape; (d) an object with different geometry. The transfered modes and residuals are combined to form the final results (bottom row).*

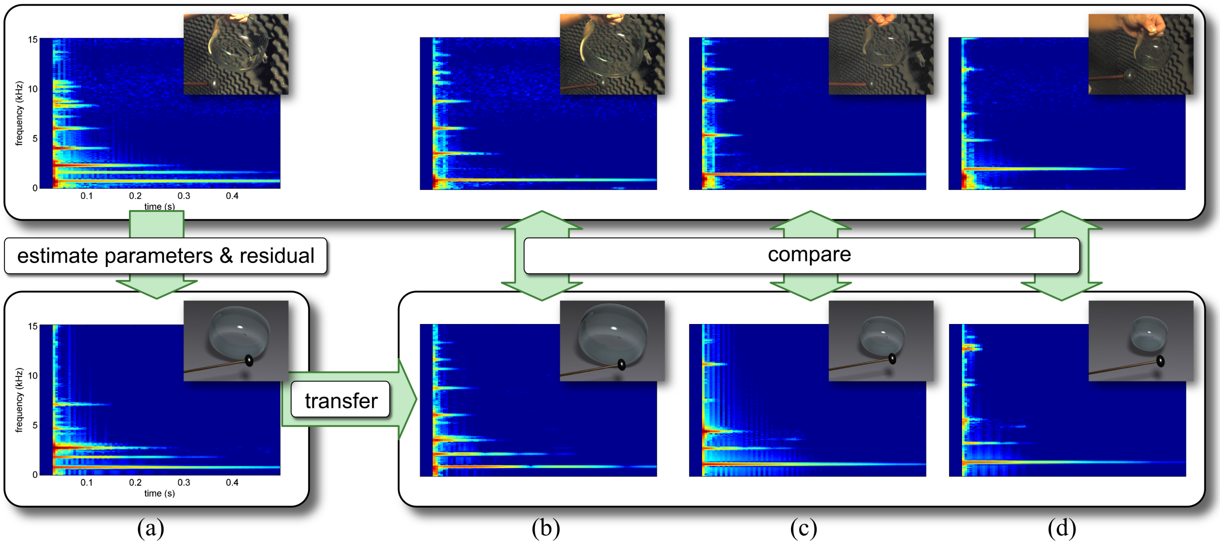


Figure 8: *Comparison of transfered results with real-word recordings: from one recording (column (a), top), the optimal parameters and residual are estimated, and a similar sound is reproduced (column (a), bottom). The parameters and residual can then be applied to different objects of the same material ((b), (c), (d), bottom), and the results are comparable to the real-world recordings ((b), (c), (d), top).*