REALIMPACT: A Dataset of Impact Sound Fields for Real Objects (Appendix)

Samuel Clarke[1] Ruohan Gao[1] Mason Wang[1] Mark Rau[1]
Julia Xu[1] Jui-Hsien Wang[2] Doug L. James[1] Jiajun Wu[1]
[1]Stanford University [2]Adobe Research

The supplementary materials consist of:

A. Supplementary video.
B. Details of the recording environment.
C. Details of the recording apparatus.
D. Additional details and observations on the nature of the hammer impacts in our dataset.
E. Additional results on interpolating from transfer maps.
F. Additional results from measuring the repeatability of measurements from objects of different materials.
G. Additional details on our simulation baselines and their assumptions.
H. Additional examples of denoised clips from our dataset.
I. Example inputs and outputs from our visual acoustic matching task.

A. Supplementary Video

In the supplementary video, we show 1) a visualization of the diverse set of objects we use, 2) illustrations and demos of our custom hardware setup and data collection process, 3) examples of the impact sound field in our dataset, and 4) comparisons of audio clips from our dataset against a series of simulation methods [2, 3, 8].

B. Recording Environment

In order to validate the recording environment and the efficacy of its acoustic treatments in reducing reverberations, we measured the room impulse response with the following procedure. We positioned a loudspeaker at the same location of the room at which we had positioned our objects during our recordings. We then played a ten-second logarithmic sinusoidal sweep from 20 - 20 kHz through the loudspeaker and recorded it with the microphone array. The gantry moved the microphone array through the same positions at which we had collected the object recordings for the dataset, and we recorded the sweep from each position. In this way, we could capture the specific impulse response at every potential recording position to ensure there was good uniformity of the environment and the recordings across all measurement positions we had used for the dataset. We converted each microphone recording of the sine sweep to an impulse response using deconvolution, and then calculated the octave-band reverberation time for the sound to decay by 60 dB (RT60) using the Schroeder method [6].

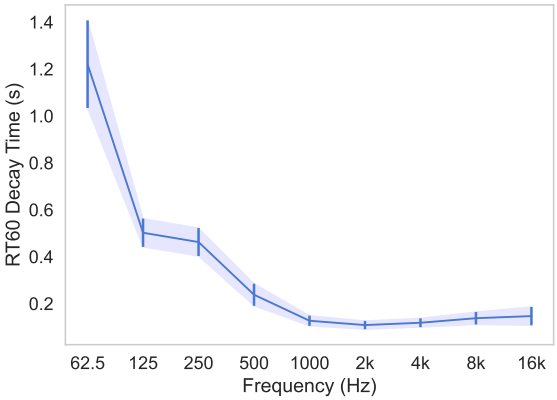

Figure 7. Octave-band RT60 measurements made in the measurement room averaged across all 600 microphone locations.

The octave band T60 measurements are shown in Figure 7. The T60 is below 0.2 s for frequencies above 500 Hz, suggesting that the room is fairly anechoic. Below 500 Hz, there is a longer reverb time, as we had made a compromise to treat the room down to a reasonable frequency range while maintaining usable space. With regards to the dataset, since most objects are small, few will have low-frequency resonant modes. Most modes are above 500 Hz, the range in which the room is least reverberant.

C. Recording Apparatus Details

Here we provide more details and explanations of the mechanical design of our recording apparatus.

Figure 8 shows the motion of the hammer striking mech-

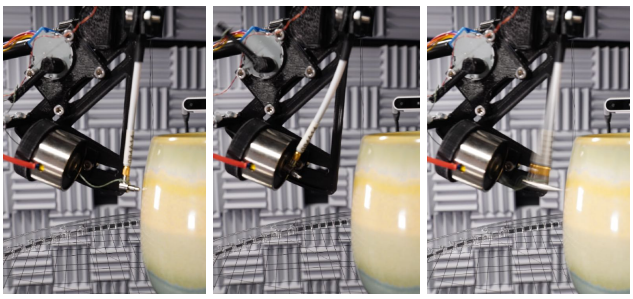

Figure 8. The automated hammer striking mechanism in action. **(Left)** We manually position the head of the impact hammer such that it is initially near the target impact point without making contact with the object at rest. **(Center)** To strike the object with the hammer, the motor first winds back the hammer, until it contacts the activated electromagnet to be held into place. The motor then unwinds while the electromagnet holds the hammer. **(Right)** Finally, the electromagnet releases the hammer to strike the object with as little noise from motion as possible.

anism. The hammer is cantilevered to the striking apparatus by the end of its handle. The handle consists of a light plastic tube with enough elasticity to store spring energy as the head of the hammer is pulled back to the electromagnet. Furthermore, because this handle is light, the inertia of the system is low enough to mitigate the risk of the head of the hammer bouncing off the object multiple times and polluting our recordings. To further ensure that we do not capture multiple hits in our recordings, we also programmatically validate the recorded signal after each recording.

Our gantry's motion elements are shown in Figure 9. The base of the gantry essentially consists of a linear slide resting on a Vention turntable located at the center of rotation, with passive fixed caster wheels at the other end. A stepper motor with a built-in encoder drives the rotation of the turntable. The stepper motor and encoder system have 800 pulses per rotation, and the turntable has a gearbox with a 10:1 gear ratio, meaning that the gantry can theoretically be controlled to 0.045° precision. However, due to the rated backlash of the turntable, the nonzero flexibility of the linear slide and gantry chassis, and the unevenness of the carpet in the room, the gantry may settle into a position up to 1° from where it has been programmed to be for a recording. The linear slide is driven by a separate identical stepper motor and encoder system, with a timing belt moving 150 mm per rotation. With 800 pulses per rotation from the stepper motor, this can theoretically be controlled to 0.19mm precision. The linear slide is more stable at rest and not as susceptible to the unevenness of the carpet in settling to a different position when the motors have been turned off. Because of the flexibility of the column and the mounts of the microphones, we estimate that effective precision of the linear motion of the gantry is 1 mm.

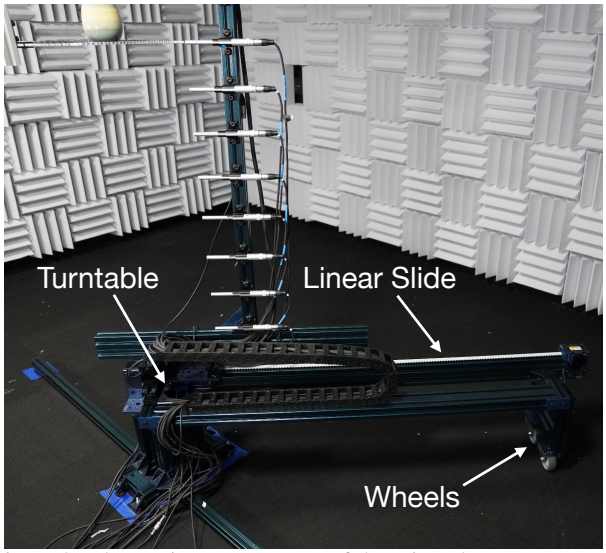

Figure 9. The motion components of the microphone gantry. For rotational motion, a stepper motor rotates a turntable at the center of the gantry, while passive wheels support the other end of the gantry and follow a circular path on the floor. For linear motion, a separate stepper motor drives a linear slide with a timing belt to precisely position the column of microphones.

D. Hammer Impacts

Our using a custom apparatus to strike objects with our steel-tipped impact hammer is important for measuring the contact forces as well as increasing precision and repeatability. We discuss some other implications of these design choices.

Impact locations We choose five striking locations for each object manually, based on multiple trade-offs. First, we generally choose striking locations which optimize for coverage of the different salient regions of each object (*e.g.*, choosing a location on the handle of a mug as well as on the side near its lip). Second, we avoid choosing two striking locations which are symmetric to each other about a plane or axis of symmetry in the geometry of the object. Third, we choose points which are *reachable* by the tip of the hammer, given that the striking apparatus limits the tip's reach. And finally, we choose points which eliminate or minimize the striking apparatus' occlusion of the line of sight between the object and any of the microphones.

Impact forces Though the striking apparatus provides a rather precise and repeatable swinging motion to the hammer, we observe some variation in the striking forces we measured for each object, object vertex, and even vertex trial. The hammer's instantaneous peak striking force is mostly a function of the hardness and restitution of each object's material, ranging from 1.07 to 298 N across the dataset, with a mean of 109 N across all objects. The average standard deviation of the peak forces across all vertices

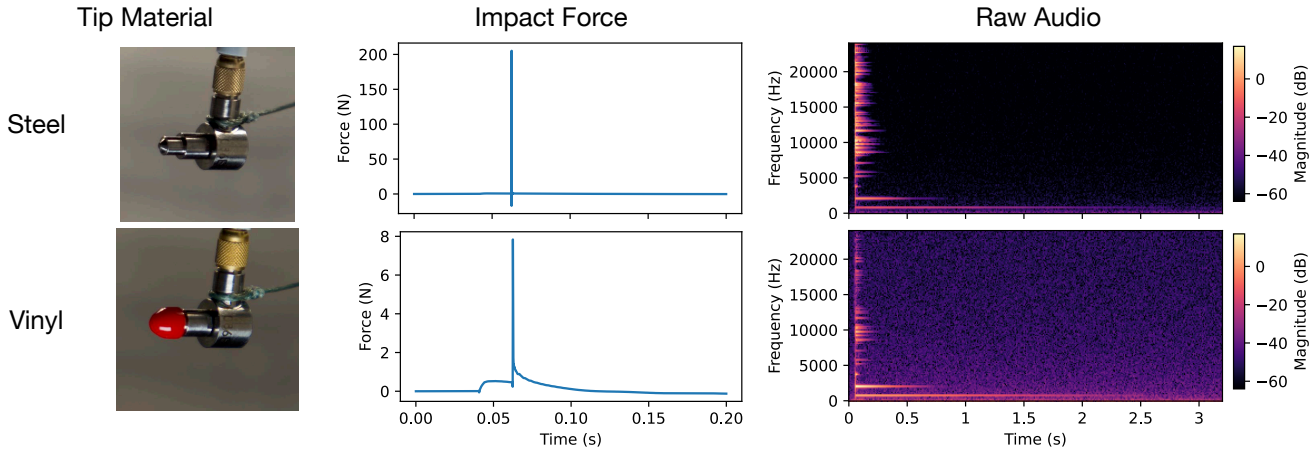

Figure 10. Comparing the resulting forces and audio of striking the ceramic bowl with different materials of impact hammer tip. The top row shows the results of using the standard steel tip as we used in our dataset. The bottom row shows the results of using the tip covered by the soft vinyl cap shown covering the steel tip of the hammer in the image in the bottom row of the left column.

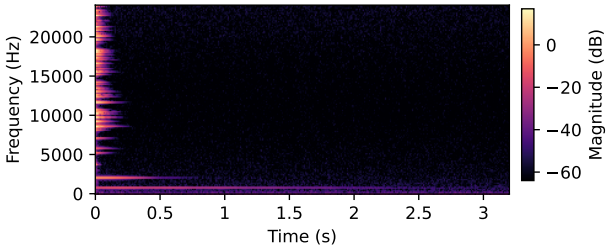

Figure 11. The resulting impulse response estimated by deconvolving the hammer contact forces from the recording of the steel tip striking the ceramic bowl shown in Figure 10.

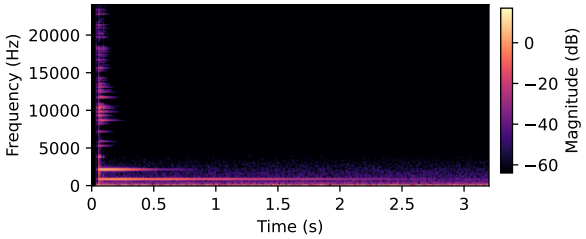

Figure 12. The result of naïvely estimating the sound of striking the ceramic bowl with the vinyl tip by convolving the impulse response from Figure 11 with the impact forces of the vinyl tip shown in the bottom of the middle column in Figure 10.

from a single object is 29.8 N, and the average standard deviation across all trials of the same vertex is 11.7 N.

Hammer material The impact hammer is comprised of a plastic handle and a hardened steel tip. The plastic handle emits minimal, but non-negligible, sound as it swings and strikes objects. The tip of the hammer is small enough that its modes of vibration all have frequencies above the Nyquist frequency of our recordings as well as human audible frequencies, thus not directly influencing our impact recordings. The hardened steel tip of the hammer maximimizes the repeatability of impacts and also ensures that impacts are as sharp as possible to both excite the high-frequency modes of each object and make each strike as loud as possible to boost the signal-to-noise ratio of our recordings. This in turn allows us to characterize the impulse response as precisely as possible. Using a softer material for the hammer creates contacts which are longer, which essentially low-pass filters the impulse response of the object [4], and softer, which decreases the signal-to-noise ratio of recordings. In order to demonstrate this, we compare the results of striking the ceramic bowl with the steel tip we used for our dataset, compared to those of striking the bowl with the steel tip covered by a soft vinyl cap in Figure 10.

Note that for the vinyl-capped tip, the duration of the impact force is indeed longer, and its peak magnitude is much lower. The audio of the impact sound from the vinyl tip is accordingly much quieter, with much more evident noise in the spectrogram confirming a lower signal-to-noise ratio.

However, we can use the deconvolved impulse response from the measurements of impacts using the steel tip to predict the sounds an object would make under different contact conditions, including being struck by a different material. The recording of the steel tip striking the ceramic bowl shown in the top row of Figure 10 yields the deconvolved impulse response shown in Figure 11. We can then convolve this impulse response with new hammer contact forces to make a naïve prediction of the sound the object would make when acted upon by those contact forces. For example, we can use this principle to predict the sound of the ceramic bowl being struck by the soft vinyl tip. We convolve the deconvolved impulse response from the steel tip with the contact forces from the vinyl tip, with the result shown in Figure 12. When compared to the ground truth audio recorded from the impact of the vinyl tip (shown in the spectrogram at the bottom right of Figure 10), the prediction shows a modal response with very similar characteristics to

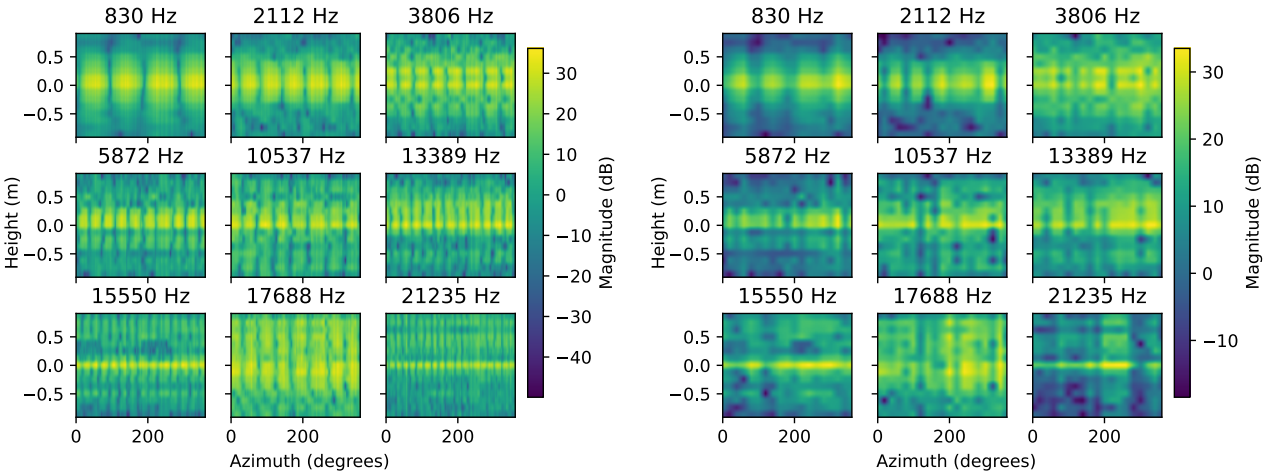

Figure 13. Comparing ground truth measurements versus interpolated mode shape transfer maps of the nine most salient modes of a ceramic bowl. **(Left)** Ground truth measurements, measured at an azimuth angle resolution of 1°. **(Right)** The results of linearly interpolating a 1° azimuth resolution from the 20° resolution measurements used in our data collection process.

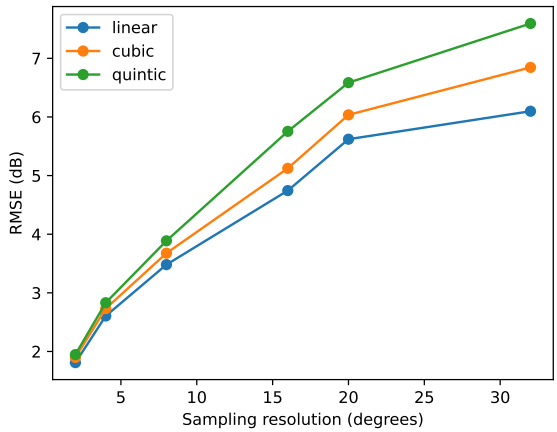

Figure 14. Comparing different interpolation methods for their error in interpolating 1° transfer maps of the ceramic bowl from different levels of azimuth angle coarseness, averaged across the bowl's nine most salient modes.

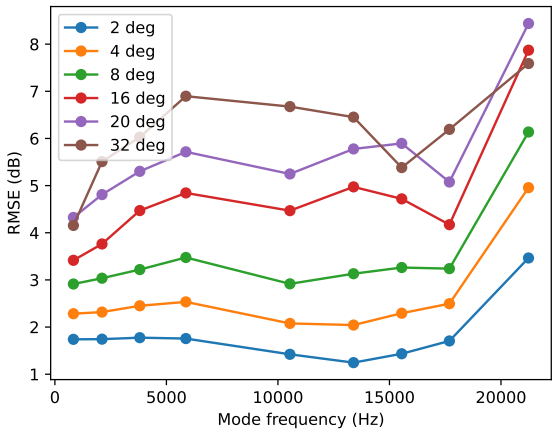

Figure 15. Error of linear interpolation toward estimating transfer maps of 1° azimuth resolution from the ceramic bowl at each mode frequency, stratified by the coarseness of azimuth angle resolution on which each interpolation is based.

that of the ground truth, yet markedly different from the modal response of the steel tip at the top left of Figure 10. Further, by using the impulse response from the steel tip with a much higher signal-to-noise ratio, the prediction is less polluted by measurement noise than the actual ground truth recording.

E. Interpolating from Transfer Maps

Here we show some results from attempting to naïvely interpolate high-resolution mode shape transfer maps from lower-resolution maps. First, in addition to those already shown in Figure 4, the ground-truth high-resolution transfer maps from salient modes of the ceramic bowl are shown on left side of Figure 13. These additional transfer maps have also been collected by the same procedure described in § 4.5 and processed by the procedure described in § 4.4.

We downsample each of these maps to increasingly coarse azimuth angle resolutions and attempt to interpolate back to 1° azimuth resolution using linear, cubic, and quintic interpolation methods, then measure the RMSE in decibels of each method at each coarseness of resolution. We average the error of each method across the mode transfer maps of each the nine frequencies and show the results in Figure 14. A simple linear interpolation outperforms the cubic and quintic interpolations at every level of coarseness. Figure 15 shows the error of linear interpolation from each coarseness of azimuth angle resolution, with separate error for each mode frequency. The mode shape transfer maps from the 13389, 15550, and 21234 Hz modes suffer the highest errors as the coarseness of sampling increases. As seen on the left of Figure 13, the transfer maps for each of these modes have especially high frequency of repeti-

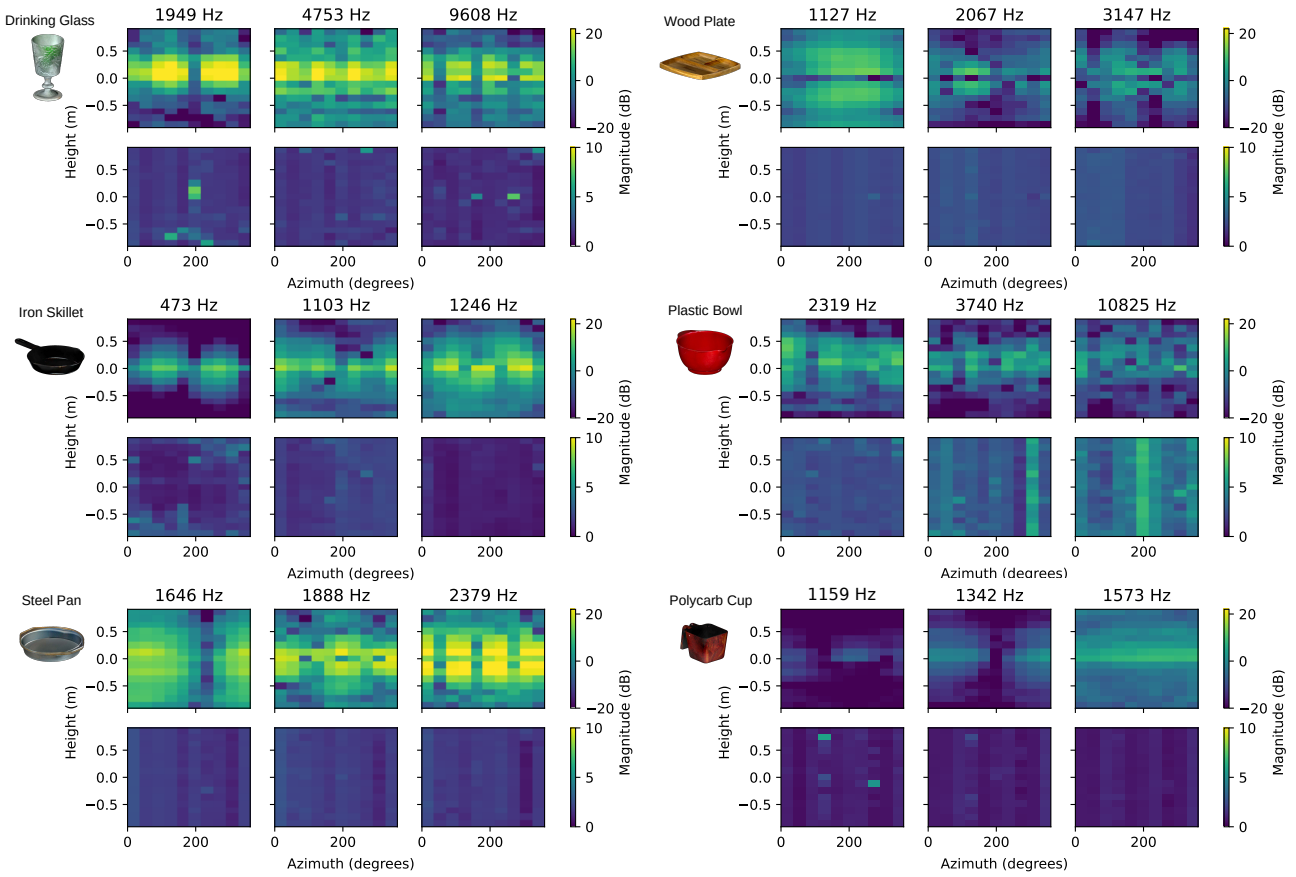

Figure 16. Measuring repeatability of our measurements by visualizing transfer maps of vibrational frequencies of the objects of different materials, measured at 23 cm from the center of each object. The top row of transfer maps for each object shows the mean of 10 trials of measurements of striking the same vertex on the object, while the bottom shows the relative standard deviation of the 10 trials.

tion of their nodes with respect to azimuth angle. Increasing the coarseness of the spatial sampling resolution beyond the Nyquist frequencies of each of these patterns is bound to expand the error of interpolation.

As described in § 4.3, our dataset provides sound fields measured at 20° azimuth angle resolution. To demonstrate the challenges of using our dataset to interpolate a sound field, we show the results of linearly interpolating 1° azimuth resolution transfer maps of the ceramic bowl from 20° transfer maps on the right side of Figure 13. These results reflect that a naïve interpolation, without any domain-specific model bias or priors, will be prone to high errors when trying to interpolate sound fields from the azimuth resolutions at which we have sampled them from our dataset. This motivates future work which is able to use priors to fit high-resolution sound fields from the spatial resolution of the sound fields in our dataset, or perhaps interpolate from an even more minimal amount of measurements.

F. Additional Repeatability Results

Along with measuring the repeatability of the ceramic bowl (Figure 5), we measured the repeatability of an object from each of the six additional materials according to the same procedure described in § 4.5, conducting 10 trials of our measurements striking a single vertex on each object. We show the mean and standard deviations of the transfers we measured at some sample modal vibrational frequencies for each object in Figures 16.

G. Baseline Details and Assumptions

As stated in § 5.1, each baseline we evaluated used different assumptions and techniques for simulating sounds. Additional details of the differences in assumptions and methods are detailed below and summarized in Table 4.

Baseline Modal Models Each baseline estimates the structural vibrations of objects through finite element-based modal analysis. NEURALSOUND computes modal analysis by voxelizing objects into hexahedral meshes, whereas KLEINPAT and ObjectFolder tetrahedralize objects to capture fine geometric features. Both NEURALSOUND and KLEINPAT use first order mesh elements, while ObjectFolder uses second order tetrahedra to model the curva-

	KLEINPAT [8]	NEURALSOUND [3]	OBJECTFOLDER 2.0 [2]
Modal Analysis & Model			
Finite Element Shape	Tetrahedral	Hexahedral	Tetrahedral
Finite Element Order	First	First	Second
Inference	Precomputed Table	LOBPCG Optimization w/ Neural Warm Start	Implicit Neural Representation
Acoustic Transfer Model			
Ground Truth Source	Boundary Element Method	Boundary Element Method	N/A
Inference	Precomputed FFAT Map	Neurally Predicted FFAT Map	N/A

Table 4. Comparing assumptions and methods of each baseline model.

ture of finite elements during modal vibrations At inference time, KLEINPAT estimates the vibrations of the object by directly computing a modal response from the frequencies and gains (*i.e.*, displacements for each mode shape) at each vertex of the mesh from the results of the LU decomposition. NEURALSOUND trains a sparse U-Net to output vectors which are used as input to the Rayleigh-Ritz method to approximate eigenvalues and eigenvectors. At inference time, the approximated eigenvalues and eigenvectors are quickly optimized using a Locally Optimal Block Preconditioned Conjugate Gradient (LOBPCG) optimization to arrive at the final eigenvalue and eigenvector estimates. ObjectFolder uses the eigenvectors estimated by Abaqus [7] to train an implicit neural representation to estimate the modal gains at any contact point on the object. At inference time, the modal response is constructed by using the frequencies estimated by Abaqus and gains predicted by the implicit representation. All baselines use the Rayleigh damping method for estimating the dampings of each mode, based on the same parameters for each material.

Baseline Acoustic Transfer Models While ObjectFolder does not model acoustic transfer, KLEINPAT and NEURALSOUND each use different methods for estimating the acoustic transfer of each object. KLEINPAT precomputes Far-Field Acoustic Transfer (FFAT) maps from performing *mode conflation* and computing transfer using a finite-difference time-domain (FDTD) wavesolver. NEURALSOUND computes FFAT maps using a Boundary Element Method (BEM) solver and uses these maps to train a ResNet-like encoder-decoder network to predict FFAT maps for each mode, using the objects' mesh and the mode frequency as input. At inference time, KLEINPAT merely uses its precomputed FFAT maps of each mode of an object, while NEURALSOUND uses its network to predict the FFAT maps to estimate acoustic transfer of each mode.

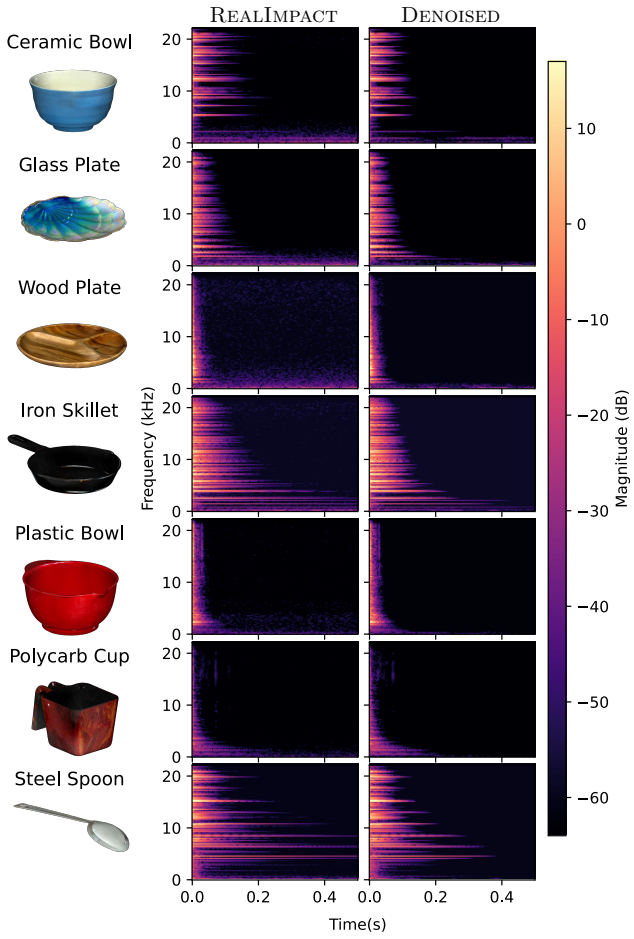

Figure 17. Example spectrograms from REALIMPACT's raw deconvolved recordings compared to their denoised counterparts, for objects of different materials.

H. Additional Denoising Examples

Additional example spectrograms of REALIMPACT's recordings compared to their denoised versions, produced by the algorithm of [5] are shown in Figure 17. The denoising algorithm seems to be especially helpful in removing the

low frequency noise for each object. This is especially evident in the recordings for the ceramic bowl, the glass plate, and the wood plate.

However, while filtering out noise, the algorithm also seems to filter out some important signal. The algorithm filters out modes after they have partially decayed, increasing their effective decay rate. Note in Figure 17 that the modal vibrations of the iron skillet and especially the steel spoon are shortened significantly in their duration by this algorithm. By effectively accelerating the decay of these modes, characterizing the objects' vibrations from these denoised versions could lead to overestimates of the damping properties of the objects and their materials. This motivates future work for an efficient denoising algorithm which is specialized for impact sounds, perhaps inspired by the physics-based principles of modal vibrations, similar to the denoising technique presented in [1].

I. Visual Matching Examples

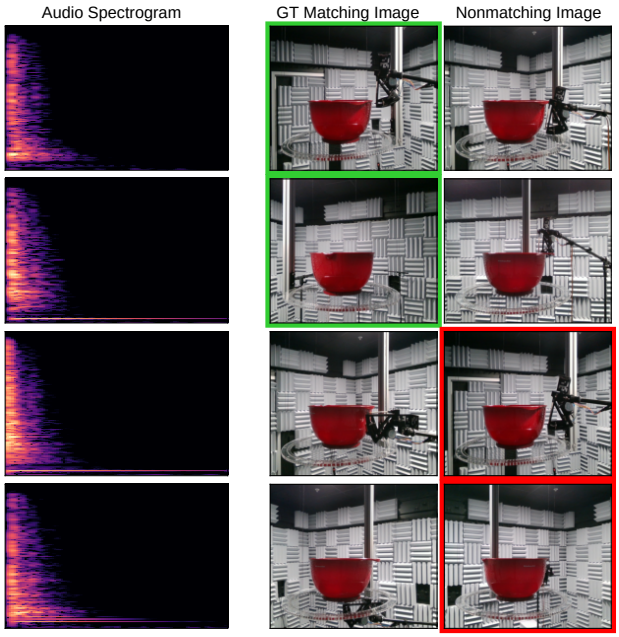

Figure 18. Example success (top two rows) and failure (bottom two rows) cases of our model for the visual acoustic matching task on a *plastic mixing bowl*.

Figures 18 - 20 show a random selection of examples of two success and two failures for three different objects in the visual acoustic matching task described in § 5.2 of the main manuscript.

For the results of the wooden wine glass shown in Figure 19, in the success cases, the different position and angle of the hammer stand and object lead to greater visual contrast between the correct matching and nonmatching images in each pair. In both failure cases, the images in each pair

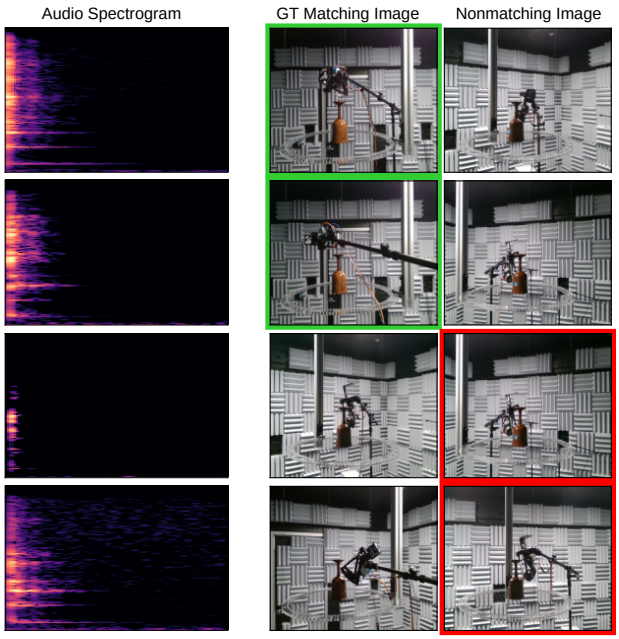

Figure 19. Example success (top two rows) and failure (bottom two rows) cases of our model for the visual acoustic matching task on a *wood wine glass*.

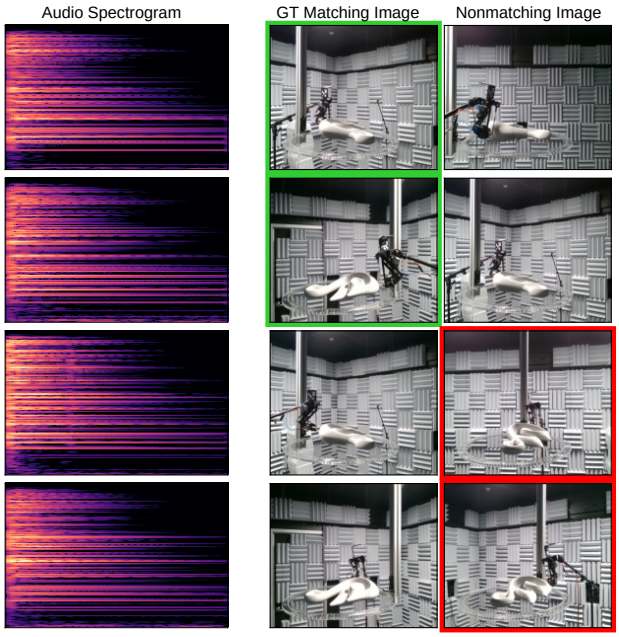

Figure 20. Example success (top two rows) and failure (bottom two rows) cases of our model for the visual matching task on a *decorative ceramic swan*.

appear to be more visually similar to each other. The hammer stand and object are located and angled in similar positions. This contrast in visual similarities and differences between success and failure cases is also evident in the re-

sults from the other objects. One important confounding factor is that the model could be exploiting and learning from the visual differences in the room. Each image also captures background details of the recording apparatus and recording room, such as the microphone stand position and patterns in the acoustic padding. It is unclear if the model is learning from the positions of the object and hammer or from other environmental factors in the room. In real-world settings, such external factors may be especially wise to exploit, since they are also likely to influence the acoustic environment of the object and therefore its sound field.

References

[1] Samuel Clarke, Negin Heravi, Mark Rau, Ruohan Gao, Jiajun Wu, Doug James, and Jeannette Bohg. DiffImpact: Differentiable Rendering and Identification of Impact Sounds. In *CoRL*, 2021. 7

[2] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. ObjectFolder 2.0: A Multisensory Object Dataset for Sim2Real Transfer. In *CVPR*, 2022. 1, 6

[3] Xutong Jin, Sheng Li, Guoping Wang, and Dinesh Manocha. Neuralsound: learning-based modal sound synthesis with acoustic transfer. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022. 1, 6

[4] KD Mali and PM Singru. Study on the effect of the impact location and the type of hammer tip on the frequency response function (FRF) in experimental modal analysis of rectangular plates. In *IOP Conference Series: Materials Science and Engineering*, 2018. 3

[5] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Computational Biology*, 16(10):e1008228, 2020. 6

[6] Manfred R Schroeder. New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(6):1187–1188, 1965. 1

[7] Dassault systemes. Abaqus. *Simulia Corporation*, 2021. 6

[8] Jui-Hsien Wang and Doug L James. Kleinpat: Optimal mode conflation for time-domain precomputation of acoustic transfer. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 1, 6