Hearing the Room Through the Shape of the Drum: Modal-Guided Sound Recovery from Multi-Point Surface Vibrations: Supplementary materials

Shai Bagon Matan Kichler Mark Sheinin
Weizmann Institute of Science, Israel
shai.bagon@weizmann.ac.il, matankic@gmail.com, mark.sheinin@weizmann.ac.il

A. Spatially varying optical transfer and mode shape estimation

In Sec. 3.2 of the main manuscript, we simplified the relationship between the measured speckle shifts $\mathbf{v}(\mathbf{x}_n, t)$ and the surface gradients $\nabla_{\mathbf{x}} u(\mathbf{x}_n, t)$ by assuming a global optical transfer factor β. However, in a general imaging scenario, effects such as a high variance between the points' axial camera distances may yield different per-point scalers, β_n. In such circumstances, Eq. (7) becomes:

$$\mathbf{v}(\mathbf{x}_n, t) = \beta_n \nabla_{\mathbf{x}} u(\mathbf{x}_n, t). \tag{17}$$

Thus, substituting the modal expansion (Eq. (6)) into Eq. (17) now yields:

$$\begin{aligned}\mathbf{v}(\mathbf{x}_n, t) &= \gamma\beta_n \sum_{k=1}^{K} \nabla\phi_k(\mathbf{x}_n)\left(s(t) * g_k(t)\right) + \eta(\mathbf{x}_n, t) \\ &= \gamma \sum_{k=1}^{K} \nabla\phi_k^{\text{aug}}(\mathbf{x}_n)\left(s(t) * g_k(t)\right) + \eta(\mathbf{x}_n, t),\end{aligned} \tag{18}$$

where we define the *augmented* mode shape gradient as $\nabla\phi_k^{\text{aug}}(\mathbf{x}_n) := \beta_n \nabla\phi_k(\mathbf{x}_n)$.

Crucially, our procedure does not require explicitly decoupling the varying factors β_n from the true physical mode shape gradients $\nabla\phi_k(\mathbf{x}_n)$. Because our mode shape extraction is completely data-driven, it inherently absorbs this per-point scaling factor. Specifically, our numerical recovery of the mode shape gradients is extracted directly using Eq. (11) (reproduced below):

$$\nabla\hat{\phi}_k(\mathbf{x}_n) = \text{Re}\left\{\frac{\mathbf{V}(\mathbf{x}_n, \hat{\omega}_k) \cdot \mathbf{V}_1(\mathbf{x}_0, \hat{\omega}_k)^*}{\mathbb{E}_{n,a}[|\mathbf{V}(\mathbf{x}_n, \hat{\omega}_k)|] \cdot |\mathbf{V}_1(\mathbf{x}_0, \hat{\omega}_k)|}\right\}. \tag{11}$$

Thus, in the case of non-negligible β_n, Eq. (11) directly approximates the augmented term $\nabla\phi_k^{\text{aug}}(\mathbf{x}_n)$ up to a global normalization constant. The factors β_n will simply skew the numerically recovered mode shape gradients by a constant per-point factor. Consequently, when we invert the model to solve for the recovered latent sound source $\hat{s}^{\text{inv}}(t)$, the optimization (Eq. (12)) remains completely agnostic to the true physical mode shape gradients of the surface:

$$\underset{s(t),\alpha_k}{\text{argmin}} \left\| \mathbf{v}(\mathbf{x}_n, t) - \sum_{k=1}^{K} \nabla\hat{\phi}_k^{\text{aug}}(\mathbf{x}_n)\left(s(t) * \hat{g}_k(t)\right) \right\|_2^2 + \lambda \left\|\dot{s}(t)\right\|_2^2. \tag{12}$$

Note that because the latent sound $s(t)$ and the modal impulse responses $\hat{g}_k(t)$ are convolved in our forward model, there is an inherent scale ambiguity. Any remaining global scaling discrepancies between the measured and physical domains are cleanly absorbed jointly by the optimized modal coupling coefficients α_k (implicit within $\hat{g}_k(t)$) and the overall amplitude of the recovered signal $\hat{s}^{\text{inv}}(t)$. Since $s(t)$ is a dimensionless audio signal, this global scale factor simply alters the absolute playback volume without affecting the spectral content or acoustic fidelity. Thus, the precise physical calibration of β_n is bypassed without compromising the robust recovery of the acoustic signal.

B. Full experimental results

We extensively evaluated our method on a variety of objects with diverse materials (wood, metal, plastic, rubber), geometries (planar, curved), and irregular shapes (such as the wooden binder and guitar body). The approach even works for solid objects like the Yoga block despite not being optimized for volumetric structures. Figure A shows all the objects along with the reconstructed spectrograms.

Quantitative evaluation We further evaluated the recovered audio using several established numerical scores. Note that we do not have a ground-truth target audio, only the reference source signal played by a speaker in front of each object. Hence, we quantitatively compare the recovered audio to the input source signal. We compare to three baselines: (i) *Single*: the raw vibrations measured at one point

ViSQOLAudio-NSIM↑				
Experiment	Single	Avg	DnS	Ours
drum	0.20	0.19	0.18	**0.32**
frame	0.22	0.15	0.23	**0.46**
laptop	0.28	0.22	0.33	**0.44**
trash	0.19	0.24	0.24	**0.43**
guitar	0.16	0.26	0.24	**0.32**
binder	0.26	0.27	0.37	**0.43**
plate	0.19	0.17	0.23	**0.34**
drum (stereo)	0.26	0.26	0.27	**0.38**
yoga block	0.16	0.08	0.16	**0.29**
physio ball	0.35	0.35	0.39	**0.39**
balloon	0.27	0.17	0.39	**0.46**
Mean:	0.23	0.21	0.27	**0.39**

Table A. **ViSQOLAudio-NSIM** [2, 3] (higher is better). The raw perceptual similarity index used inside ViSQOLAudio before Mean Opinion Score (MOS) mapping. It measures structural similarity between cochleagram patches of the reference and degraded signals and is not tied to speech-specific MOS training. Higher values indicate closer perceptual similarity.

ViSQOLAudio-MOS↑				
Experiment	Single	Avg	DnS	Ours
drum	**2.05**	1.90	1.96	1.63
frame	1.22	1.16	1.31	**1.96**
laptop	1.44	1.61	1.90	**2.06**
trash	1.07	1.12	1.05	**2.58**
guitar	**1.74**	1.59	1.66	1.58
binder	1.47	1.54	1.55	**2.45**
plate	1.24	1.43	1.45	**1.95**
drum (stereo)	1.69	1.65	1.55	**1.84**
yoga block	1.16	1.00	1.19	**2.67**
physio ball	1.37	1.54	1.32	**1.57**
balloon	1.58	1.43	**1.86**	1.81
Mean:	1.46	1.45	1.53	**2.01**

Table B. **ViSQOLAudio-MOS** [3] (higher isbetter). Perceptual audio-quality metric based on the ViSQOLAudio model. It compares a reference and degraded signal using a cochleagram front-end and patch-based similarity, then maps the result to a Mean Opinion Score (MOS) prediction. Originally trained on general-audio listening tests, it approximates subjective quality.

on the surface, (ii) *Avg*: naive averaging of all measured vibrations, and (iii) *DnS*: delay-and-sum where a single global temporal delay is estimated per point [1].

Table A reports ViSQOLAudio-NSIM [2, 3], which is the raw perceptual similarity index used inside ViSQOLAudio before mean opinion score (MOS) mapping. Higher ViSQOLAudio-NSIM score indicates better similarity between the recovered audio and the input source signal. Our recoveries consistently score higher than all baselines. We further used the mean opinion score (MOS) mapping on top of ViSQOLAudio (Tab. B).

Inspired by [4, 5], We further used scale-invariant multi-

Scale Invariant, Multi Resolution STFT distance↓				
Experiment	Single	Avg	DnS	Ours
drum	4.68	4.21	4.60	**3.54**
frame	3.02	3.46	3.15	**1.87**
laptop	4.01	5.17	4.59	**3.88**
trash	3.65	3.27	3.46	**3.01**
guitar	3.23	**2.47**	2.59	3.49
binder	3.00	3.51	3.14	**2.72**
plate	3.22	3.60	3.82	**3.21**
drum (stereo)	**4.27**	4.56	4.41	4.32
yoga block	3.92	4.20	**3.66**	3.68
physio ball	**2.92**	3.40	3.16	4.00
balloon	**3.18**	3.39	3.96	3.73
Mean:	3.55	3.75	3.69	**3.40**

Table C. **Scale Invariant, Multi Resolution STFT distance** [4, 5] (lower is better). A distance metric between recovered signal and reference source signal. The measure is a scale invariant, multi resolution STFT distance.

Scale Invariant, Multi Resolution STFT distance (Perceptual weighting)↓				
Experiment	Single	Avg	DnS	Ours
drum	3.65	3.39	**3.28**	3.39
frame	3.64	5.01	3.94	**2.01**
laptop	3.11	3.66	3.92	**2.91**
trash	3.56	3.57	3.88	**2.61**
guitar	3.45	2.91	3.04	**2.53**
binder	3.00	3.18	3.20	**2.48**
plate	3.44	3.57	4.37	**2.70**
drum (stereo)	3.67	**3.59**	4.16	3.97
yoga block	4.14	4.01	3.87	**3.68**
physio ball	**3.08**	3.22	3.12	3.35
balloon	**3.00**	3.12	3.42	3.58
Mean:	3.43	3.57	3.65	**3.02**

Table D. **Scale Invariant, Multi Resolution STFT distance (Perceptual weighting)** [4, 5] (lower is better). A distance metric between recovered signal and reference source signal. The measure is a scale invariant, multi resolution STFT distance with perceptual weighting.

resolution distance computed in the space of short-time Fourier Transform (STFT). Table C shows the basic distance, while Tab. D shows a perceptually-weighed distance computed in this space. On average, our results have smaller distance to the reference input signal compared to all other baselines quantitatively showing the superiority of our audio recovery.

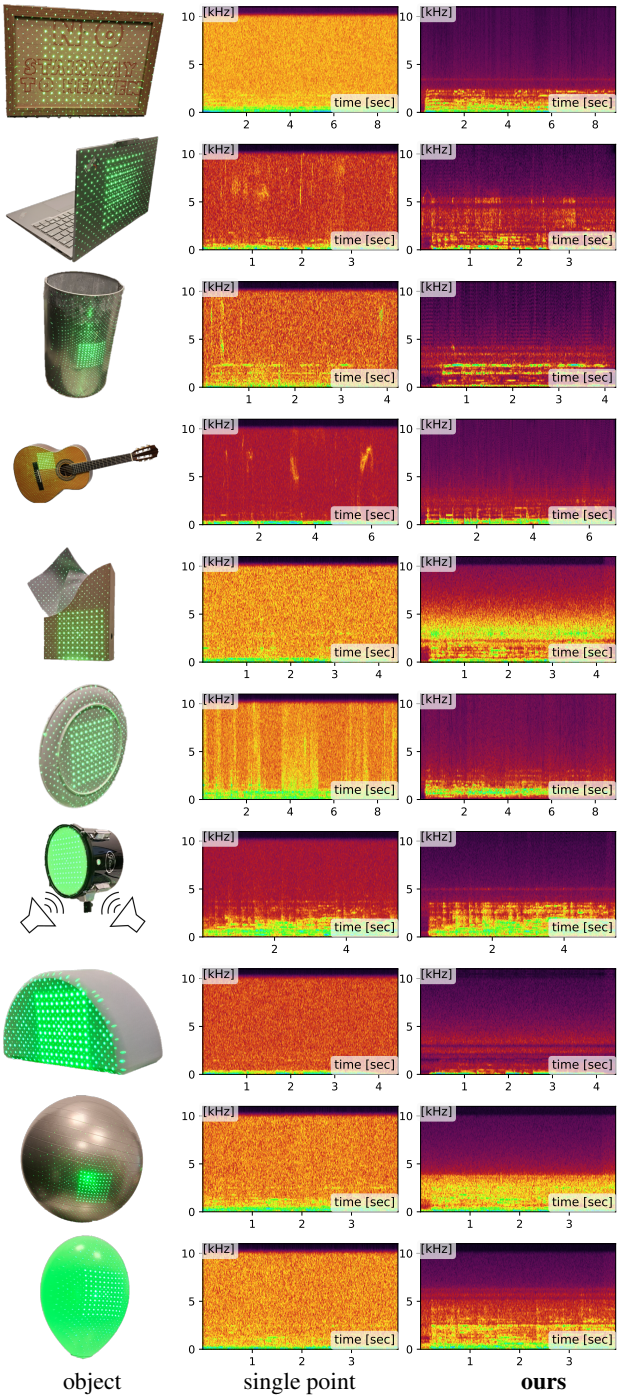

Figure A. Results across objects having various geometries and materials. All objects present challenges due to diverse materials, thicknesses, and shapes. Despite limited or irregular surface coverage, our method yields denoised reconstructions.

References

[1] Jacob Benesty, Israel Cohen, and Jingdong Chen. *Fundamentals of Signal Enhancement and Array Signal Processing*. John Wiley & Sons Singapore Pte. Ltd., 2017. 2

[2] Andrew Hines and Naomi Harte. Speech intelligibility prediction using a neurogram similarity index measure. *Speech Communication*, 54(2):306–320, 2012. 2

[3] Andrew Hines, Eoin Gillen, Damien Kelly, Jan Skoglund, Anil Kokaram, and Naomi Harte. ViSQOLAudio: An objective audio quality metric for low bitrate codecs. *The Journal of the Acoustical Society of America*, 137(6):EL449–EL455, 2015. 2

[4] Christian J. Steinmetz and Joshua D. Reiss. auraloss: Audio focused loss functions in PyTorch. In *Digital Music Research Network One-day Workshop (DMRN+15)*, 2020. 2

[5] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. 2