Event2Audio: Supplementary Materials

Mingxuan Cai*, Dekel Galor*, Amit Pal Singh Kohli, Jacob L. Yates, and Laura Waller

1 Additional Experimental Results

1.1 Motion Dynamic Range

The motion dynamic range is important when capturing subtle vibrations. Unlike conventional cameras, event cameras have configurable motion sensitivity via a custom contrast threshold, such that weak light changes are filtered out as noise. To evaluate the system's response across different motion ranges, we conducted an experiment in which the system's behavior was assessed as a function of shift size (which we approximated as the input audio amplitude).

Specifically, we played an up-chirp signal with amplitude ranging from 0.1 to 1.0 in increments of 0.1. To measure how the system's performance degrades, we compute the maximum reconstructible frequency for the default threshold value. Before the experiment, we reduced the speaker volume to calibrate the system such that the maximum reconstructible frequency was approximately 2.3 kHz when playing the chirp signal at an amplitude of 1.0.

In Fig. 1(top), the maximum reconstructible frequency decreases progressively with lower amplitudes (e.g., 1954 Hz at amplitude 0.8; 695 Hz at amplitude 0.3). At an amplitude of 0.1, no discernible signal was observed in the reconstruction, indicating a failure to capture the vibration under this condition. In parallel, we computed the number of events recorded at each amplitude level in Fig. 1(bottom), which exhibited a logarithmic growth trend from 0.1 to 1.0. Notably, the trend of the maximum reconstructible frequency roughly followed that of the event count.

Therefore, a sufficient large motion or carefully adjusted motion threshold is needed for high-quality event-based vibration sensing. In our study, our imaging system has the ability to amplify subtle motion by adjusting the focus, which may help mitigate this issue.

1.2 Large Motion

We conducted a proof-of-concept experiment to evaluate the system's performance under large motion. We played a 440 Hz signal (Fig. 2(b)) while hand-holding the speaker to induce large scene motion. The induced motion displacement reaches ∼2000 pixels in the X dimension and ∼3000 pixels in the Y dimension (Fig. 2(a)), which is comparable to the large motion reported in [1].

When zooming in on the motion trajectory in Fig. 2(a), the waveform corresponding to the 440 Hz signal becomes clearly visible. Consequently, in the reconstruction shown in Fig. 2(c), we successfully recovered the 440 Hz audio despite the presence of substantial scene motion.

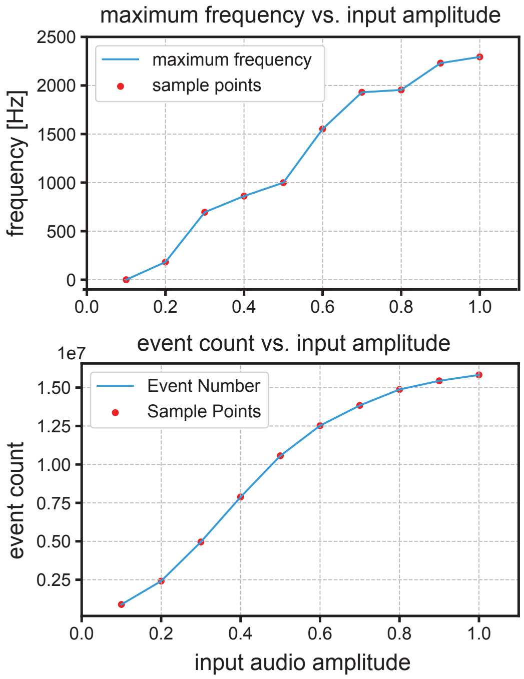

Fig. 1. Experimental results of motion dynamic range evaluation. **(Top)** The maximum reconstructible frequency with different input audio amplitudes from 0.1 to 1.0. **(Bottom)** The recorded event count with different input audio amplitudes from 0.1 to 1.0.

1.3 Multiple Laser Spots

We conducted an additional experiment involving more than two laser spots and audio sources. Specifically, we implemented our system with three laser spots and corresponding audio sources. Consistent with our two-source experiment, speckle patterns from different sources were projected onto distinct regions of the sensor, enabling physical separation of the signals.

In this setup, we played different tones corresponding to different pitch classes through three separate speakers (Figs. 3(b–d)) while simultaneously recording the mixed audio with a microphone. As shown in Fig. 3(a), the microphone captured a complex superposition of the three audio signals. In contrast, our system successfully isolated and reconstructed the individual audio signals, as shown by the accurate and clearly resolved chromagrams in Figs. 3(e–g).

Other than audio source separation, Zhang *et al.* [2] demonstrated that having more laser spots enables the analysis of surface vibration waves. This in turned allowed

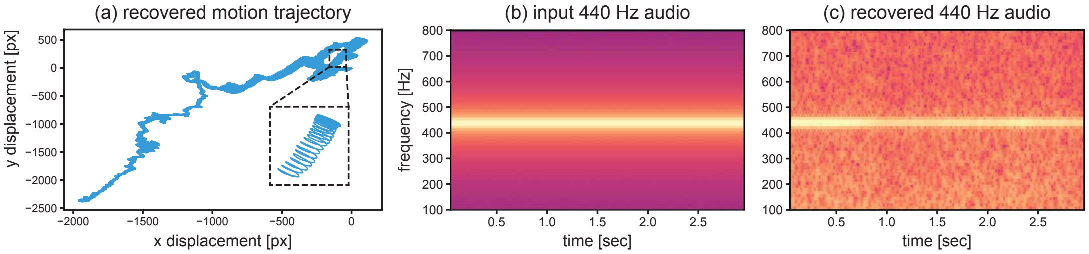

Fig. 2. Experiment results with large hand-holding motion. **(a)** Recovered motion trajectory. **(b)** Input 440 Hz audio. **(c)** Recovered 440 Hz audio.

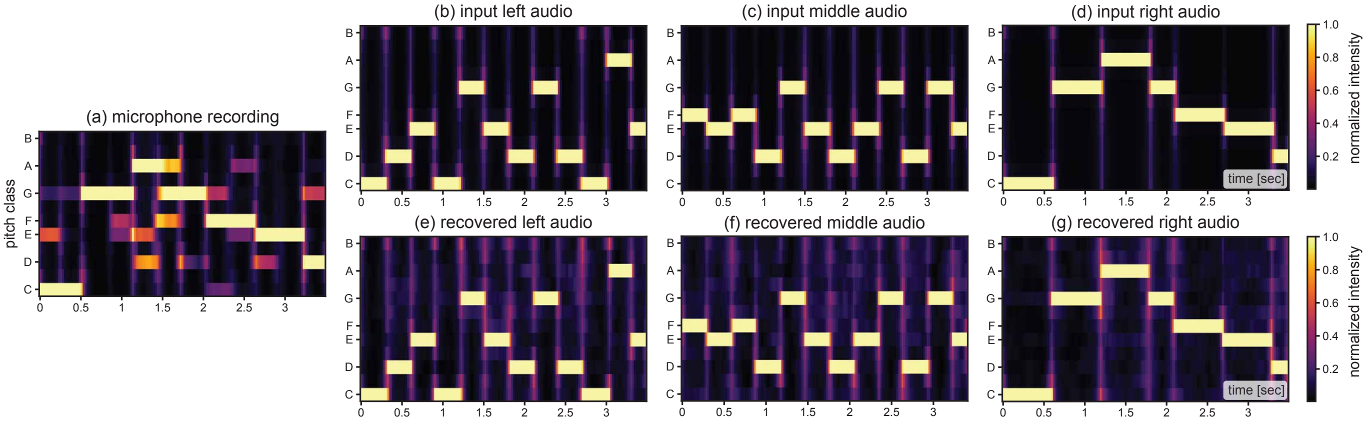

Fig. 3. Capturing signals from three laser spots and audio sources. **(a)** The chromagram of the microphone recording. **(b-d)** The chromagrams of the input audio from left, middle, and right speakers. **(e-g)** The chromagrams of the recovered left, middle, and right audio signals.

for impact localization and analysis of material properties, which are exciting avenues for future exploration of event-based vibrometry.

References

[1] M. Sheinin, D. Chan, M. O'Toole, and S. G. Narasimhan, "Dual-shutter optical vibration sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 324–16 333.

[2] T. Zhang, M. Sheinin, D. Chan, M. Rau, M. O'Toole, and S. G. Narasimhan, "Analyzing physical impacts using transient surface wave imaging," in *Proc. IEEE CVPR*, 2023.