# Event-based Visual Vibrometry

## Supplementary Material

Xinyu Zhou[1] Peiqi Duan[2,3] Yeliduosi Xiaokaiti[2,3] Chao Xu[1] Boxin Shi[2,3*]

[1]State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University

[2] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

[3]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

{zhouxiny, duanqi0001, shiboxin}@pku.edu.cn, yongqiye@stu.pku.edu.cn, xuchao@cis.pku.edu.cn

## 6. Additional experimental results

This section presents additional experimental results demonstrating the efficacy of our method in visual vibrometry applications.

### 6.1. Audio recovery under different lighting

The high dynamic range (HDR) characteristic of event cameras enables event-based visual vibrometry to operate effectively under ambient illumination conditions. To analyze the performance of our method under varying illumination levels, we record one speaker playing a chirp signal at four distinct brightness levels, ranging from 400 lux to 3200 lux. For comparison of the frame-based method, we simultaneously capture the speaker with a high-speed video camera at 1000 fps and recover audio signals using the visual microphone (VM) technique [2]. The signal reconstruction quality is evaluated using the segmental signal-to-noise ratio (SSNR). As shown in Tab. 4, our method achieves more robust performance across different lighting conditions. Notably, the performance of our method under low illumination (400 lux) is comparable to that of the frame-based method under high illumination (3200 lux).

Table 4. Signal reconstruction SSNR (the higher the better) comparison under different lighting conditions.

| Method | 400 lux | 800 lux | 1600 lux | 3200 lux |
|---|---|---|---|---|
| VM [2] | 0.31 | 0.38 | 0.89 | 2.25 |
| Ours | 3.75 | 4.48 | 4.51 | 4.65 |

### 6.2. Analyzing vibration of tuning forks

We analyze the vibrations of two tuning forks with fundamental frequencies of 128 Hz and 256 Hz, respectively. As shown in Fig. 7, we strike the forks with a rubber-tipped mallet and measure their vibrations. The spectrograms obtained by our method accurately reflect the fundamental frequencies of the tuning forks. In contrast, EBVM fails to recover these fundamental frequencies, potentially due to noise in the event signal under ambient lighting conditions.

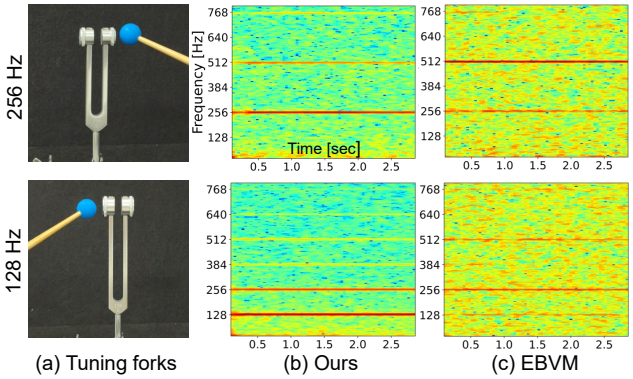


Figure 7. Vibration analysis of two tuning forks with fundamental frequencies of 128Hz and 256Hz (a). We compare recovered spectrograms between our method (b) and EBVM [5] (c). The results obtained using our method align well with the fundamental frequencies of the tuning forks.

### 6.3. Material properties with unknown geometry

We demonstrate the applicability of our method in learning the material properties of objects with unknown geometry. The experiments on material property estimation, as detailed in the main manuscript, rely on precise knowledge of the object's geometry. As a result, their potential application is limited to objects with simple geometries that can be precisely measured, or to man-made structures with detailed CAD models, for which resonant frequencies can be obtained through the finite element method (FEM).

Given a set of objects with similar but not precisely modeled geometries, the differences in their material properties will be revealed in their resonant frequencies and mode shapes. Based on this intuition, Davis *et al*. propose to learn relationships between motion spectra and the material properties of objects with similar but unknown geometry. They conducted experiments on a dataset comprising 30 hanging fabrics [1], along with corresponding ground truth measurements of area weight. Following their work, we simu-

* Corresponding author

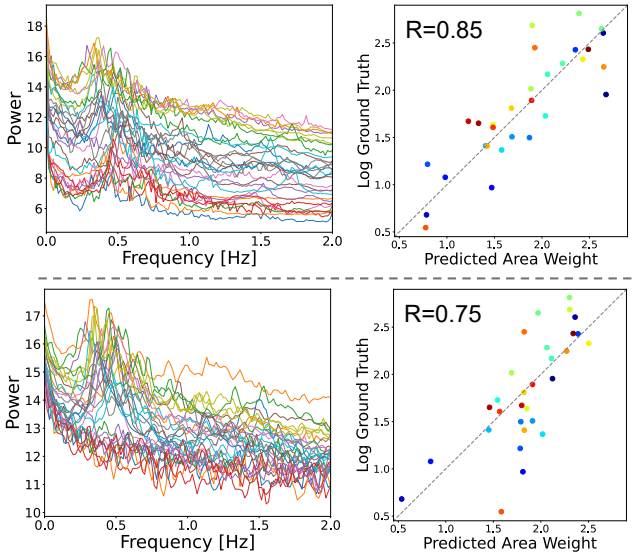


Figure 8. Comparison of extracted motion spectra and predictions on material properties estimated from videos [3] (upper) and events (below) on the fabric dataset [1]. The Pearson correlation values (R) are shown in the figure.

late corresponding events with the V2E simulator [4] from videos captured by a grayscale Point Grey camera at 60 fps. From the simulated events, we extract motion spectra using our proposed method. Consistent with the methodology in [3], we employ the motion spectra directly as features and train a Partial Least Squares Regression model to map the motion spectra to the logarithm of the ground truth area weight. Due to the small size of the dataset, we employ the leave-one-out cross-validation strategy. The extracted motion spectra and area weight prediction results obtained from our method and from videos are presented in Fig. 8. The Pearson correlation values (R) of our predictions are slightly lower than those of the frame-based method. Note that the data size of our simulated events (16-bit Prophesee EVT 3.0 format) is less than 10% that of the original videos, indicating that the vibrations are efficiently encoded in the simulated events. Considering the reduced data size, we believe the minor performance drop is acceptable.

# 7. Discussion

## 7.1. Thresholds' impact on vibration sensing

For frame-based cameras, the accuracy of vibration estimation depends on bit depth and quantum efficiency in noise-free conditions. Correspondingly, the precision of subtle motion estimation from event data is affected by the contrast threshold. The threshold of event cameras generally exceeds one gray level in images, resulting in a lower theoretical precision for event-based visual vibrometry, especially in scenarios involving extremely low-amplitude vibration measurements. Intuitively, lowering the threshold would trigger more event signals, thereby improving motion extraction accuracy. Nevertheless, due to current hardware constraints, event cameras are more susceptible to noise at lower thresholds. Despite this hardware limitation, experimental results demonstrate that our method achieves satisfactory performance across many applications. Future hardware advancements will further improve the precision of event-based visual vibrometry.

## 7.2. Inference frequency

The time step for voxel partitioning is primarily determined by the vibration frequency of the observed object. Specifically, the Nyquist frequency of the motion estimation results should exceed the target frequency range. The computational overhead of our method increases linearly with the number of voxels. At a resolution of 256×256, the overall running time for each step of the coarse motion optimization and subsequent network refinement on our test system is approximately 0.04s.

We evaluate the quality of audio recovered from the sequence capturing a speaker playing the "MarySpeech" audio file used by Davis *et al*. [2] under 4 distinct inference frequencies: 1000, 2000, 4000, 6000Hz. The intelligibility (STOI) scores are [0.481, 0.513, 0.524, 0.523] (higher is better). Increasing the inference frequency expands the detectable vibration spectrum, thereby enhancing signal reconstruction fidelity. However, this concurrently reduces the event signal density per voxel, which may affect the precision of micro-vibration estimation. Correspondingly, the STOI scores exhibit an initial ascent followed by a plateau. In our experiment, we set the inference frequency slightly above the target frequency range.

## 7.3. Temporal filtering

Previous motion magnification studies typically employ temporal filtering to select motion within specific frequency bands of interest using a band-pass filter. Temporal filtering helps to prevent noise from being magnified, but it requires prior knowledge of the observed vibration. In contrast, our method usually aims to analyze vibrations across a broad frequency spectrum. To mitigate noise in subtle motion estimation, our approach employs a reference image and exploits the temporal structure inherent in event data, which is extracted through a recurrent event encoder, thereby effectively suppressing isolated noise events.

## 7.4. Dynamic scenes

When the observed object undergoes global motion, it is more challenging to detect subtle vibrations. Following previous studies [2, 3], our method assumes that the observed object remains static, exhibiting only tiny vibrations. Under this assumption, our approach utilizes only one im-

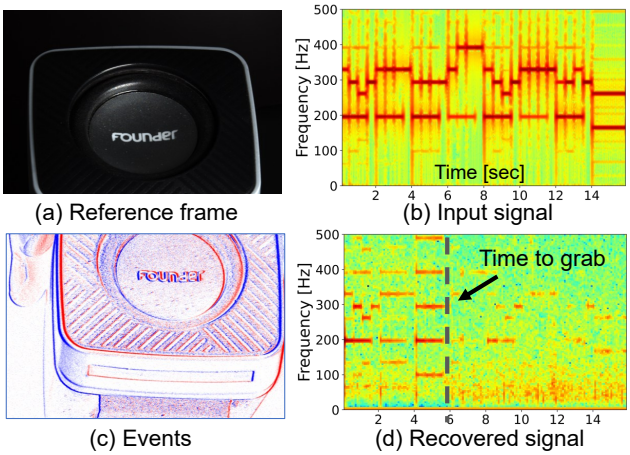


Figure 9. We capture a speaker that is manually grasped and shaken. (a) The reference frame. (b) The spectrogram of the input signal sent to the speaker. (c) Event signals at another timestamp. (d) The spectrogram of our recovered sound.

age to provide scene texture information and is inapplicable to dynamic scenes. To evaluate our method on non-static scenes, we conduct an experiment where a speaker is manually grasped and shaken. As shown in Fig. 9, the spectrogram of the recovered signal reveals a performance drop when the speaker is shaken. Future works could track the object's macro-motion using both frames and events, and dynamically update the reference frame.

## References

[1] Katherine L Bouman, Bei Xiao, Peter Battaglia, and William T Freeman. Estimating the material properties of fabric from video. In *Proc. of IEEE International Conference on Computer Vision*, 2013. 1, 2

[2] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J. Mysore, Frédo Durand, and William T. Freeman. The visual microphone: passive recovery of sound from video. *ACM Transactions on Graphics*, 33(4):79:1–79:10, 2014. 1, 2

[3] Abe Davis, Katherine L Bouman, Justin G Chen, Michael Rubinstein, Oral Büyüköztürk, Frédo Durand, and William T Freeman. Visual Vibrometry: Estimating material properties from small motions in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):732–745, 2017. 2

[4] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbrück. V2E: From video frames to realistic DVS events. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[5] Ryogo Niwa, Tatsuki Fushimi, Kenta Yamamoto, and Yoichi Ochiai. Live demonstration: Event-based visual microphone. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1