Event2Audio: Event-Based Optical Vibration Sensing

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

Abstract—Small vibrations observed in video can unveil information beyond what is visual, such as sound and material properties. It is possible to passively record these vibrations when they are visually perceptible, or actively amplify their visual contribution with a laser beam when they are not perceptible. In this paper, we improve upon the active sensing approach by leveraging event-based cameras, which are designed to efficiently capture fast motion. We demonstrate our method experimentally by recovering audio from vibrations, even for multiple simultaneous sources, and in the presence of environmental distortions. Our approach matches the state-of-the-art reconstruction quality at much faster speeds, approaching real-time processing.

Index Terms—Optical Vibration Sensing; Vibrometry; Event Cameras;

1 Introduction

The imperceptible vibrations in everyday objects contain rich information about their composition and the environment they interact with. Optical vibrometry enables remote measurement of such vibrations using visual sensors. This makes optical vibrometry an important tool across various engineering disciplines, where it is often used for fault detection [1], [2], sound recovery [3], and inferring material properties [4], [5].

When employing such systems for dynamic tasks like fault detection or sound recovery, vibration signals must be reconstructed quickly and reliably from optical measurements, thus creating a need for robust, high-speed optical vibrometry [6]. In this work, we take a critical step towards this ultimate goal by developing a new vibrometer using event-based sensing [7] that approaches real-time processing speeds while maintaining high-quality reconstruction.

Existing optical vibrometers trade off robustness, sensitivity, speed, and design simplicity. Passive vibrometers directly image the surface of a vibrating target [4], [5], [8], allowing for simple hardware at the cost of sensitivity—they can only measure motion that is visually observable [9]. On the other hand, active vibrometers [10], [11] illuminate the target and measure the motion signal from reflected light, offering increased sensitivity but at the expense of robustness and simplicity [1]. Recent work uses active modalities that encode the reflected light with simple hardware but require sophisticated, slow post-processing algorithms to unearth the vibration signal [9], [11]. The key is striking a balance where the vibration is optically encoded using simple hardware while still allowing for fast decoding—something we achieve by taking advantage of the unique properties of event-based cameras.

These cameras [7], [12] record changes in log-intensity as asynchronous events, each carrying a timestamp, spatial location, and polarity (i.e., the sign of the intensity change). This sensing paradigm enables extremely high temporal resolution while maintaining low memory and processing overhead [13]. The sparse and information-rich event streams that come from the event camera provide an ideal solution for co-optimizing performance and speed [14], since they reduce redundant information (i.e., static background pixels), which only serve to slow down post-processing algorithms. Inspired by their natural fit for real-time sensing, we incorporate them into a novel active vibrometry system. Our method reflects coherent light off of the target object, creating a motion-sensitive speckle image at the event sensor. The event sensor only records the aspects of the speckle that move with the object vibrations, allowing us to create a near real-time post-processing algorithm that maps events to the underlying vibration signal.

We demonstrate our method experimentally by recovering audio from vibrations, even in the presence of multiple simultaneous sources and environmental distortions. Our approach matches the state-of-the-art (SOTA) reconstruction quality at much faster speeds approaching real-time processing. We make the following contributions:

- We present a novel event-based vibrometry method for recovering audio signals from vibrating surfaces. The proposed system is simple, compact, and does not require complex alignment.
- Leveraging this novel system, we introduce an audio reconstruction algorithm approaching real-time processing with significantly better performance than other near real-time methods.
- We introduce a second offline algorithm that matches SOTA performance with an order-of-magnitude speedup over existing SOTA methods.
- We validate our method on audio signals in varying experimental conditions, outperforming traditional microphones in challenging environments.

2 Related Work

2.1 Optical Vibrometry

Optical vibrometry is generally categorized into two types: passive and active sensing. The distinction depends on whether active illumination, such as a laser, is used.

- *M. Cai and D. Galor contributed equally to this work.*
- *M. Cai, D. Galor, A. P. Kholi, L. Waller are with the Department of Electrical Engineering and Computer Sciences, J. L. Yates is with the Herbert Wertheim School of Optometry and Vision Science, University of California, Berkeley, Berkeley, CA, 94720.*
- *E-mail: {mingxuan_cai, galor, apkohli, yates, waller}@berkeley.edu*

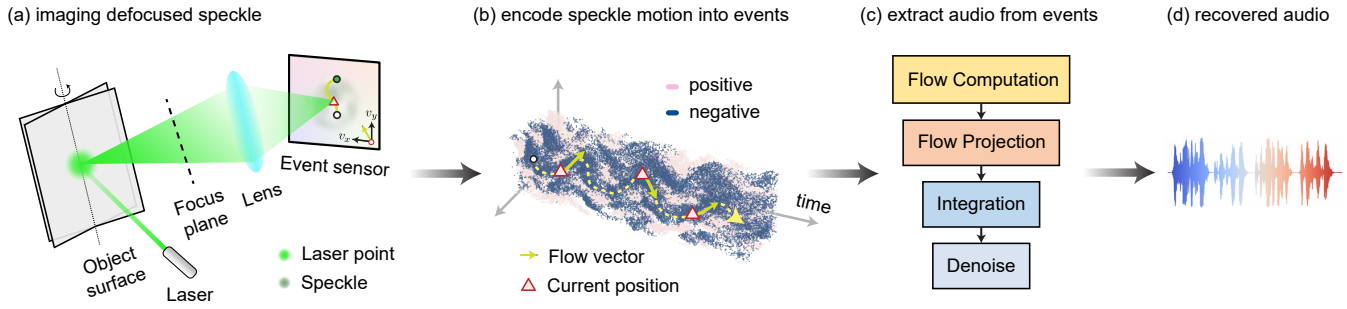

Fig. 1. Schematic of the proposed method. **(a)** Imaging defocused speckle. A coherent laser illuminates the vibrating surface, generating a defocused speckle pattern on the sensor plane. The pattern's 2D movements are captured by the event sensor. **(b)** The captured motion is encoded into a stream of asynchronous events. This event stream reflects the motion of the speckle pattern induced by surface vibrations. For each event, a corresponding optical flow vector, consisting of a timestamp and a 2D spatial velocity, can be extracted through optical flow computation. **(c)** Audio signal extraction from events. **(d)** Recovered audio waveform.

Passive sensing. Passive sensing methods rely on high-speed cameras to directly image vibrating surfaces [5], [15], [16], [17] affected by audio sources [8], [11], [18], [19]. To further extract subtle vibrations with low amplitude, motion magnification algorithms are often applied [20], [21], [22]. Although passive sensing is simple to implement in hardware and does not require laser sources, it has notable limitations. For example, it struggles to capture visually imperceptible vibrations, which are associated with higher frequencies in natural signals [23]. This leads to degraded reconstruction quality for higher frequencies [9], [18], even if the sensor itself is fast enough. Moreover, its effectiveness significantly diminishes when applied to long-distance audio extraction [24].

Active sensing. Modern active sensing involves illuminating the rough surface of a vibrating object with a coherent laser source, producing a speckle pattern in the far field [25], [26]. Therefore, small tilts of the vibrating surface are translated into lateral shifts in the speckle pattern [3], [27], [28], [29]. To enhance low-amplitude motions, the speckles are typically defocused before being recorded by the camera sensor [9], [27], [30]. High-speed cameras are commonly used to capture the high-frequency motion of speckles [8]. However, they are often cost-prohibitive and their improved temporal resolution comes at the expense of reduced spatial resolution [8], [9]. As an alternative, line-based cameras are much cheaper but still offer comparable temporal resolution [3], [24], [31]. They achieve this by projecting the image onto sensor rows, exposing one line at a time rather than the entire frame. Nevertheless, their reconstruction accuracy heavily depends on the alignment between the speckle motion and the sensor rows: if the motion is not parallel to the rows, the reconstructed amplitude may be significantly degraded. To address this limitation, a hybrid method combining frame-based and line-based cameras has been proposed [9], [32]. This approach leverages the speed of line-based cameras and the global capture capabilities of frame-based cameras. As a result, it enables high-fidelity reconstruction regardless of speckle motion orientation. However, this approach is complex, requiring multiple cameras, precise optical alignments, and the algorithm is still constrained by the reconstruction speed, taking hours to process a few seconds of data.

Our method implements an active sensing approach to capture imperceptible vibrations that are generally undetectable by passive sensing. Compared to previous active sensing approaches, it achieves a simple and compact optical setup without requiring multiple cameras or complex alignments. Despite this simplified design, the method delivers high-quality audio recovery comparable to state-of-the-art techniques, while maintaining much faster reconstruction speeds.

2.2 Event-based Vision

Event cameras [7], [33], also known as neuromorphic cameras, are an emerging technology for capturing fast dynamic scenes. Unlike conventional cameras that integrate light intensity over a fixed exposure time, event cameras only respond to brightness changes asynchronously and independently at each pixel at the microsecond level [12]. The unique properties of event cameras enable unmatched capabilities for capturing high-speed motion [34], making them ideal for applications such as eye tracking [35], fluid particle tracking [36], and dynamic speckle analysis [37].

Event-based vibrometry. Recent studies have explored the use of event cameras for vibrometry [18], [19]. For instance, Howard *et al.* [18] implemented a passive sensing setup using an event camera and recovered audio by analyzing the zero-crossings of pixels. This method enabled real-time audio reconstruction at an unprecedented speed. However, their approach relies solely on the timing of brightness changes. As a result, it produces binary waveforms that lack amplitude information, which affects the reconstruction quality. A similar passive sensing strategy was adopted by Niwa *et al.* [19] with a different motion extraction algorithm.

Notably, both methods are implemented with a passive setting and are limited to detecting visually perceptible vibrations, such as guitar strings or blinking LEDs. These constraints significantly reduce the effective frequency range and degrade the overall reconstruction fidelity.

In contrast, our approach leverages active sensing to optically amplify subtle vibrations, enabling high-quality reconstruction over a broader frequency range. By developing an appropriate event-based processing method, we accurately track speckle motion, resulting in superior recon-

struction quality and improved processing speed approaching real time.

3 Proposed Method

Our method consists of two main components: hardware acquisition and software processing. In the acquisition stage, we employ active illumination to encode object vibrations into the motion of a speckle pattern. An event camera is then used to efficiently capture this motion and transmit it to a computer as a stream of asynchronous events [13]. The software stage processes these events to recover the vibration signal, which is achieved by computing and integrating the optical flow.

In the following subsections, we will discuss the details behind each component of our pipeline shown in Fig. 1.

3.1 Defocused Speckle Image Formation

A *laser speckle* [38], [39] is a random intensity distribution generated when a coherent laser beam reflects off a rough surface. The reflected light comprises numerous coherent wavelets, each originating from different microscopic elements of the surface. These dephased yet coherent wavelets interfere both constructively and destructively, resulting in a random spatial interference pattern known as speckle [40].

The laser speckle is highly sensitive to small surface changes, making it an effective tool for detecting vibrations. As illustrated in Fig. 1(a), a coherent laser forms a small spot on the rough surface of the target object. When the surface vibrates, it generates a translated speckle pattern, whose motion is subsequently magnified by the defocused imaging setup before being detected by the camera [11], [30].

Although a vibrating object can exhibit multiple types of motion simultaneously, the effects of transverse and axial motion on the speckle's shape and displacement become negligible when speckle is defocused with the lens [11], [30]. Under these conditions, tilt motion becomes the dominant contributor to speckle displacement on the image sensor.

Moreover, the speckle shift remains consistent and global across the camera plane because the illuminated spot size is small, resulting in a large memory effect range [41], [42]. Within this range, small tilts of the vibrating surface—such as those induced by sound waves—do not significantly alter the speckle pattern itself. Instead, the entire speckle field undergoes the same tilt, leading to a global translation on the sensor.

As a result, the camera captures a consistent global motion of the speckle pattern that directly reflects the surface tilt induced by vibration.

3.2 Event Data

This global translation of defocused speckle is imaged with an event camera, resulting in a stream of events (Fig. 1(b)). Each event contains a timestamp t in μs, 2D pixel position (x, y), and polarity p indicating an increase ($p = 1$) or decrease ($p = -1$) in log-intensity $\log(I(x, y, t))$. A basic model of event triggering can be written as [43]:

$$(\log(I(x, y, t)) - \log(I(x, y, t_0))) \cdot p > \epsilon, \tag{1}$$

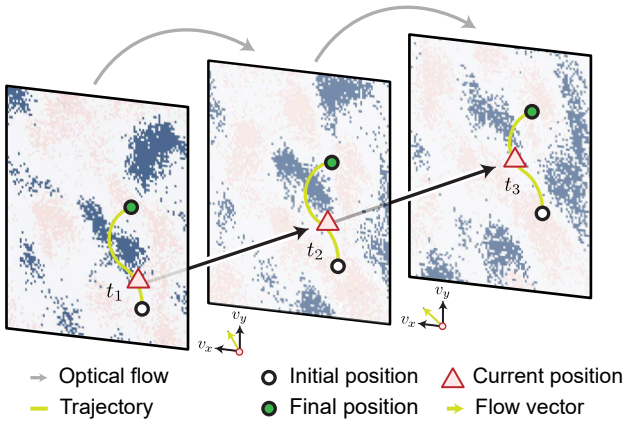

Fig. 2. Schematic of tracking speckle from events via optical flow. The flow indicates the current direction of motion, which can be integrated to find the motion trajectory. Pink: positive events. Blue: negative events.

Where t_0 the timestamp of the latest event in position (x, y), and ϵ is the contrast threshold.

The motion of the speckle, driven by surface vibrations, is thus translated into temporal and spatial patterns within the event stream. These patterns inherently encode audio information, enabling the reconstruction of sound from the captured event stream and underlying speckle dynamics.

3.3 Optical Flow Computation

At this stage, we compute optical flow from the event data to estimate global motion. This is the most computationally intensive stage of our pipeline, so it must be performed efficiently to support fast and robust vibrometry reconstruction. To accommodate different applications, we provide two alternative approaches tailored for either *real-time* or *offline* processing, depending on the desired trade-off between reconstruction speed and quality.

The *real-time* method uses a fast event-based optical flow algorithm [13] that leverages the sparse nature of events [44], [45], operating in a streaming fashion. For each event e_i, characterized by its timestamp t_i^{center}, position (x_i, y_i), and polarity p_i, the flow is estimated based on the most recent polarity-matching events in its spatial neighborhood. Specifically, for each axis, the algorithm searches for the latest events with the same polarity located r pixels away in four cardinal directions: t_i^{left} at $(x_i - r, y_i)$, t_i^{top} at $(x_i, y_i + r)$, t_i^{right} at $(x_i + r, y_i)$, and t_i^{bottom} at $(x_i, y_i - r)$. We use $r = 7$ pixels, which we found to perform well empirically in our experiments. The local flow is then computed as the spatial difference divided by the temporal difference. Notably, only the most recent matching event is retained per axis, allowing at most one neighboring horizontal and one neighboring vertical event to be used for flow estimation. For example, if the right pixel had a more recent timestamp than the left, the flow in the x direction for event e_i is computed as $\frac{\Delta x}{\Delta t} = \frac{(x_i + r) - x_i}{(t_i^{\text{right}} - t_i^{\text{center}})}$.

Each flow estimate carries a timestamp, spatial coordinates, and a velocity vector. For global motion estimation we discard the spatial coordinates and retain only the velocity components. To obtain a temporally dense representation,

we quantize timestamps to a high rate (100 kHz in our experiments), and aggregate all events that fall within the same quantized time bin, yielding a dense global flow signal with separate x and y velocity channels.

This method operates with extremely short integration windows, focusing on the temporal characteristics of motion. It offers super-fast reconstruction speed and is well-suited for applications that require live processing. However, because it ignores the broader spatial context across the object, it is more susceptible to noise and more sensitive to parameter tuning, which may impact reconstruction quality.

As an alternative, we provide an *offline* implementation. Here, events are first temporally integrated to form dense frames. We then apply the Gunnar Farnebäck algorithm [46], [47], a classical optical flow method that uses a pyramidal strategy to capture motion at multiple spatial scales. The resulting flow frames are spatially averaged (each pixel weighted by the number of events it received) to get the global flow. This multiscale approach preserves broad spatial content, enabling more robust extraction of global motion while alleviating the need for extensive parameter tuning. Unlike the *real-time* method, this integrate-first strategy does not leverage the asynchronous nature of events and is consequently slower. However, the integration makes the algorithm more robust to noise, and enables the use of mature image-based optical flow algorithms. Thus, the *offline* method acts as a benchmark and suitable alternative when live processing is not needed. Moreover, since the hardware is identical to the *real-time* method, users can seamlessly switch between the two. In practice, the *offline* method is still reasonably fast and yields more stable and accurate reconstructions.

3.4 Recovering Audio Signal from Motion

After obtaining the 2D global motion of the speckle, we convert it into a 1D audio signal by projecting the motion traces onto a single dimension while preserving essential vibration information. In practice, we perform the projection by first temporally aligning the horizontal and vertical motion calculations to account for phase differences (using cross-correlation). The two are then averaged, and the result is integrated over time to reconstruct the 1D waveform. Although different frequency components may oscillate along different directions, this projection enables us to retain the primary temporal characteristics of the motion necessary for accurate audio reconstruction.

3.5 Denoising

We further process the recovered audio signal to improve its signal-to-noise ratio. Since our reconstruction pipeline integrates motion, noise can build up over time, resulting in a gradual drift in the output signal in low frequency [8]. To mitigate this effect, we apply a high-pass Butterworth filter. The cutoff frequency is carefully chosen to preserve most meaningful audio content; in our experiments, it ranged between 10 and 100 Hz.

In addition, to further reduce background noise and enhance speech intelligibility, we used the noisereduce Python library [48], [49], which we found to be effective with almost negligible processing time. We empirically found that using

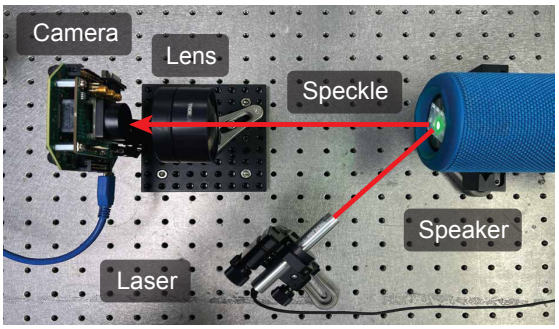

Fig. 3. Experimental system prototype. The laser illuminates the membrane of the speaker. Then, the reflected speckle pattern is captured by the lens and event camera.

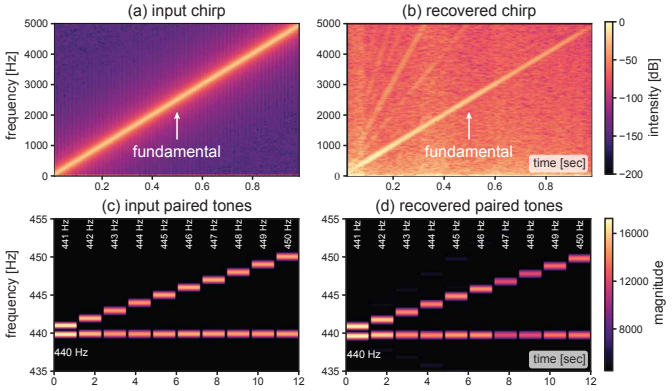

Fig. 4. Experimental results for recovering a chirp signal and paired tones. **(a)** A spectrogram of the input up-chirp signal sent to the speaker from 0 to 5 kHz. **(b)** Recovered chirp spectrogram using our method. **(c)** A spectrogram of the input paired tones. The reference tone is at 440 Hz with another tone from 441 Hz to 450 Hz. **(d)** Recovered paired tones spectrogram using our method.

80% strength, with frequency mask smoothing of 50 Hz, temporal mask smoothing of 100 ms, and a window size of 100 ms worked well across experiments and methods.

4 Prototype and Implementation Details

Hardware prototype. Figure 3 illustrates our prototype system. The system utilizes a 532 nm, 4.5 mW laser (Thorlabs CPS532) to illuminate the target surface, creating a small laser spot. Given the low power of the laser, a small patch of retro-reflective tape is affixed to the measured surfaces to enhance speckle intensity. Importantly, this modification does not affect the overall measurement results [9]. For data acquisition, the system consists of an event camera (Prophesee Metavision EVK3-HD), equipped with a 75 mm achromatic doublet lens (AC508-075-A-ML). The sensor is positioned slightly away from the focal plane of the lens to magnify the speckle pattern for subtle vibrations.

Data processing. The default bias parameters were used for the camera, and no built-in denoising was performed on the raw event data. Processing was done on a Lenovo Legion Slim 5 Laptop with an AMD Ryzen 7 7840HS CPU.

Evaluation metrics. To assess the performance of our method, we adopted evaluation metrics that are widely used in speech processing. Specifically, we employed the *Perceptual Evaluation of Speech Quality* (PESQ) [50], *Short-Time*

Objective Intelligibility (STOI/intelligibility) [51], *Mel Cepstral Distortion* (MCD) [52], and *Log Spectral Distance* (LSD) [53]. These metrics provide a quantitative assessment of speech quality, intelligibility, and spectral fidelity.

5 Experimental Results

We illustrate our method by capturing and recovering vibrations generated by various audio signals, such as chirp signals, tones, and human speech. Furthermore, we extend the utility of vibrometry to environmental distortions, including noisy, echoing, and underwater conditions, showcasing its robustness against environmental distortions and its potential for future applications. We use the built-in microphone of the iPhone 13 Pro as a reference for the octave notes and distortion experiments. Audio examples of the recovered speech are provided in the supplementary material and our website[1]. Our code and data are available on GitHub [2].

5.1 Capturing Audio Signals

5.1.1 Chirp

The most intelligible portion of human-perceived audio is typically concentrated within a frequency range up to 3.4 kHz [54]. To evaluate our system's ability to recover audio signals, we reconstructed chirp signals spanning a wide frequency range (0–5 kHz). When playing the up-chirp signal (Fig. 4(a)), the vibration of the speaker membrane induces lateral motion of the speckle on the sensor plane. As shown in Fig. 4(b), we successfully recover the fundamental frequency of the up-chirp signal. Additionally, harmonic signals emerge due to the physical properties of the speaker membrane, a phenomenon commonly observed in musical instruments such as tuning forks.

5.1.2 Frequency Resolution for Paired Tones

Frequency resolution refers to the system's ability to differentiate between closely spaced frequencies in audio signals. To assess the frequency resolution capability of our system, we played pairs of tones with frequency separations ranging from 1 Hz to 10 Hz. As illustrated in Fig. 4(c), we set a reference tone at 440 Hz while adjusting the second tone incrementally from 441 Hz to 450 Hz. As shown in Fig. 4(d), our vibrometry approach successfully distinguished signals even with 1 Hz spacing, underscoring its ability to reconstruct complex signals with densely packed spectral components.

5.1.3 Note C Octaves: Vibrometry vs. Microphone

In Fig. 5(a), we played octaves of notes C1 to C8 similar to [9]. This spans a wide frequency range from 33 Hz to 4186 Hz. The audio was also recorded using an iPhone microphone (Fig. 5(b)) and reconstructed using our vibrometry method (Fig. 5(c)). From the result, both the microphone recording and our reconstruction exhibit high-quality performance. However, consistent with the findings in [9], the microphone failed to capture the low-frequency tone at 33 Hz (white arrows in Fig. 5(right)). This is due to the low-frequency sensitivity limitations of the microphone and the built-in high-pass denoising algorithms. Interestingly, our microphone recording detected the C2 note at 65 Hz but with reduced intensity (red arrows in Fig. 5(right)), displaying a frequency response drop-off in the low-frequency region. In contrast, our vibrometry method directly measured the physical vibrations of the speaker membrane, allowing us to accurately recover both the 33 Hz and 65 Hz tones, demonstrating superior performance in low-frequency signal reconstruction.

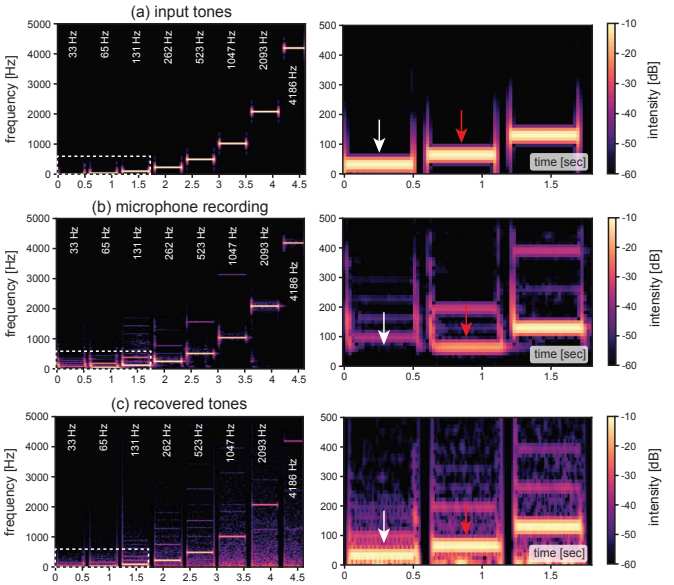

Fig. 5. Recovered octaves of the note C, from C1 (33 Hz) to C8 (4186 Hz). A single speaker plays eight octaves while a microphone records the audio. Our system captures the speaker membrane's vibrations and reconstructs the audio. The right column corresponds to a zoomed-in version of the left column. **(a)** Spectrogram of input tones. **(b)** Spectrogram of tones as recorded with a microphone. It is evident that the microphone recording fails to capture low-frequency tones due to its frequency response limitations and filtering algorithms. **(c)** Spectrogram of recovered tones. By directly sensing the physical vibrations of the speaker, our system successfully recovers these tones.

5.2 Remote Recording of Multiple Audio Sources

Demixing multiple signals from different audio sources is typically a challenging task [55], [56]. In this section, we demonstrate a configuration of our system that can simultaneously capture and separate multiple audio sources. As shown in Fig. 6(a), two different lasers illuminate the membranes of two speakers, generating spatially separated speckle patterns on the sensor plane (Fig. 6(b)). To assess the system performance, we played different tones for different pitch classes through two speakers (Fig. 6(c-d)) while simultaneously recording with a microphone. We used chromagrams to analyze audio that can be meaningfully categorized into different pitch classes. Figure 6(d) shows that the microphone captured a blended mixture of both audio signals from the left and right speakers, making it difficult to separate them directly from the recording. By leveraging the wide sensor size of the event camera, our system effectively captured vibrations from both speakers at the same time. Fig. 6(f-g) presents the recovered chromagrams from the left

1. https://mingxuancai.github.io/event2audio
2. https://github.com/dgalor/Event2Audio

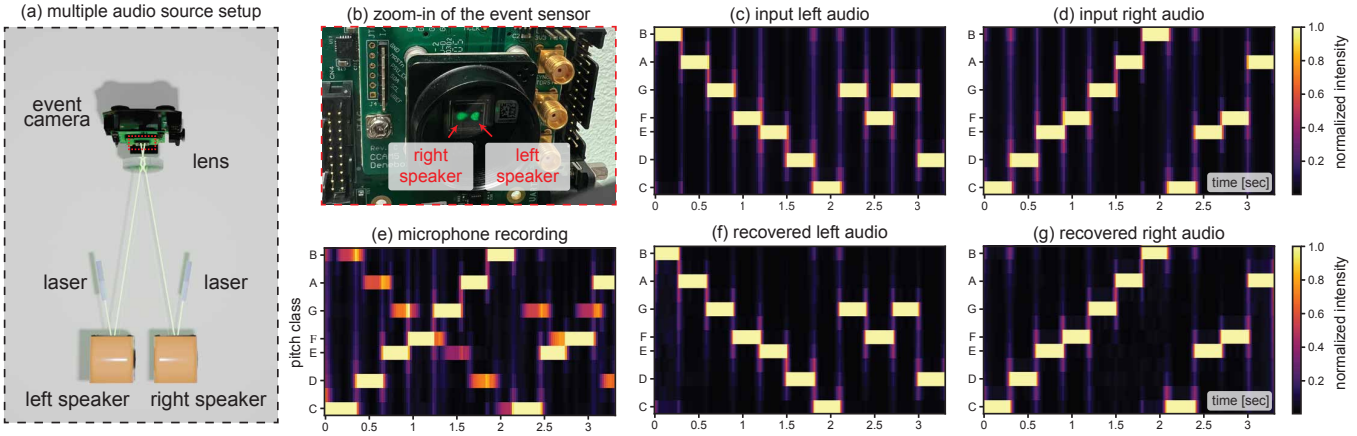

Fig. 6. Capturing signals from multiple audio sources. **(a)** Two lasers illuminate two different speakers simultaneously with a slightly different angle. **(b)** Zoomed-in view of the event sensor, corresponding to the red box in (a). The generated magnified speckle patterns are projected onto different parts of the event sensor. **(c-d)** The chromagrams of the input audio from left and right speakers. **(e)** The chromagram of the microphone recording. **(f-g)** The chromagrams of the recovered left and right audio signals.

and right speakers, which precisely match the input audio signals. This result highlights the effectiveness in separating multiple sound sources using our method. Additional experiments with more laser spots and audio sources are provided in the supplementary material.

5.3 Human Speech Recovery

5.3.1 Speech Recovery from Speaker Membrane

In addition to single-frequency tones, we extend our experiment to recover human speech signals from the vibrations of a speaker membrane. Previous studies by Niwa *et al.* [19] have achieved speech recovery from speaker vibrations but reconstruction quality still remains a bottleneck. Specifically, Niwa *et al.* attached a rod to the speaker membrane, converting its vibrations into rod oscillations. However, this approach fails to accurately capture the actual vibrations of the speaker and remains ineffective in capturing high-frequency motion. Figure 6(b) and (d) illustrate the recovered Japanese speech file "Arutoki, kita..." (Fig. 7(d)), provided by Niwa *et al*. While both the passive event camera and high-speed camera successfully recover frequency components up to approximately 1.1 kHz, they struggle to retain higher-frequency details, resulting in significant information loss. In contrast, in Fig. 7(c) and (d), our method achieves a more faithful and accurate recovery of high-frequency vibrations both in real-time and offline reconstruction. This is because our active sensing approach not only captures subtle membrane vibrations with high precision but also effectively amplifies them. Table 1 demonstrates that our method outperforms previous approaches across various evaluation metrics ([PESQ, STOI, MCD, LSD] = [1.26, 0.67, 6.72, 2.50] in real time and [1.49, 0.74, 5.57, 1.90] offline), showing a superior performance in speech reconstruction.

5.3.2 Speech Recovery from a Chip Bag

In Fig. 8(a), we replicate the chip bag experiment originally proposed by Davis *et al.* [8]. In our setup, we directed a laser onto the surface of a chip bag while playing the audio "Mary Had a Little Lamb...", as provided by Davis *et al*. In Fig. 8(b-c), we compared our real-time reconstruction results with

TABLE 1
Quantitative comparison for the experiments on recovering human speech playback from speaker membrane.

Real-time	PESQ ↑	STOI ↑	MCD ↓	LSD
Passive event sensing [19]	1.05	0.52	17.28	2.50
Ours	1.26	0.67	5.87	2.50
Offline	PESQ ↑	STOI ↑	MCD ↓	LSD ↓
High-speed camera [19]	1.16	0.45	17.87	3.50
Ours	1.49	0.74	5.57	1.90

a previous event-based vibrometry method proposed by Howard *et al.* [18]. Their approach relies on detecting zero-crossings at individual pixels and can only recover a binary amplitude waveform, resulting in degraded reconstruction quality. In contrast, our method successfully recovers the audio with high quality, which outperforms Howard's approach in all evaluation metrics in Table 2. The scores are [PESQ, STOI, MCD, LSD] = [1.04, 0.15, 26.10, 3.26] and [1.40, 0.79, 14.13, 2.41], ours being the latter.

Furthermore, other vibrometry approaches [8], [9] that achieve high-quality reconstruction typically depend on computationally intensive audio recovery algorithms, often requiring several hours to reconstruct only a few seconds of audio. In Fig. 9, we compare the reconstruction quality and processing time of our offline method with prior offline approaches. Despite different setups, which makes a direct and fair comparison difficult, it is still clear that our system demonstrates comparable reconstruction quality to the state-of-the-art method proposed by Sheinin *et al.* [9], as shown in Table 2, while significantly reducing reconstruction time by at least 30×—from approximately 54 minutes to just 1.5 minutes.

More importantly, Sheinin *et al.* employed a complex system that combines global shutter and rolling shutter cameras, along with a cylindrical lens to project speckle patterns onto the rolling shutter sensor columns. This design introduces significant challenges in optical alignment and increases hardware complexity. Additionally, splitting the

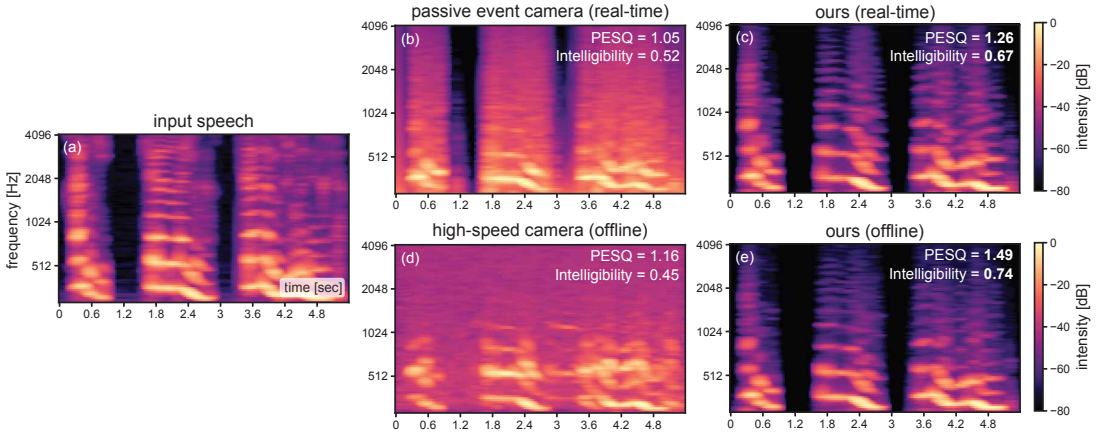

Fig. 7. Comparison of human speech recovery from speaker membrane with previous methods. A speaker played the Japanese audio file "Arutoki, kita…" provided by Niwa *et al.* [19], with baseline reconstruction results taken from that study. **(a)** Spectrogram of input speech audio. **(b)** Spectrogram of the speech audio recovered by a baseline method utilizing a passive, event-based system with real-time processing. **(c)** Spectrogram of speech audio recovered by our method with real-time processing. **(d)** Spectrogram of speech audio recovered by a method utilizing a passive, high-speed camera system with offline processing. **(e)** Spectrogram of speech audio recovered by our offline method, which provides improved reconstruction quality.

optical path between two cameras reduces light efficiency, further limiting its practicality for real-world applications.

In summary, our method features simple and compact optics, eliminating the need for complex optical alignment. Additionally, it leverages the efficient data processing capabilities of event-based sensing, enabling high-quality reconstruction with significantly faster processing on the order of minutes up to real time.

TABLE 2
Quantitative comparison for the experiments on recovering human speech playback from chip bag.

Real-time	PESQ ↑	STOI ↑	MCD ↓	LSD ↓	Time ↓
Howard [18]	1.04	0.15	26.10	3.26	real-time
Ours	1.40	0.79	14.13	2.41	real-time
Offline	PESQ ↑	STOI ↑	MCD ↓	LSD ↓	Time ↓
Davis [8]	1.06	0.44	18.31	2.16	2-3 hours
Sheinin [9]	1.39	0.73	12.79	2.20	54 min
Ours	1.60	0.81	12.56	1.72	1.5 min

5.4 Vibrometry Against Environmental Distortion

A microphone captures all sounds within its range, known as ambient sound capture, but this can introduce environmental distortions such as background noise, echoing in enclosed spaces, or attenuation underwater due to the sound waves interacting with the environment. In this section, we first propose three compelling applications where vibrometry outperforms a traditional microphone by mitigating environmental distortions. These examples demonstrate the advantages of vibrometry and highlight its potential for future applications beyond conventional audio sensing.

5.4.1 Demix Audio from Noisy Scenes

Imagine you are in a conference room, but the speaker's voice is drowned out by background noise from the audience. In such cases, you may seek a method to eliminate background noise. Noise removal from audio has been extensively studied for decades [57], [58], [59]. Existing approaches typically post-process the mixed audio signal via band filtering and, more recently, machine learning. These approaches are flexible in separating various, and potentially many, different sources from each other—even ones that were co-located during recording. However, their quality is largely limited by the difficulty of the problem and they often require massive paired datasets to achieve even moderate performance [60], [61]. In contrast, vibrometry allows for the physical isolation of the target audio source from its noisy environment at the time of recording. This means each source is essentially recorded independently, allowing for much higher quality source separation. For example, in our work, the system actively senses the surface vibrations of the target speaker, enabling direct access to its sound signal before it propagates and mixes with other sources in the environment. This contrasts with microphones, which capture a composite sound field containing both the target and background noise. However, this does come at the cost of flexibility, since the sources in question must be physically separated and there must be sufficient space on the sensor to adequately disambiguate the signal from each source.

To evaluate our approach, we simulate a noisy environment where two speakers are generating audio; one is our target speaker, and the other speaker's audio is considered background noise. As shown in Fig. 10(top), the target speaker is illuminated by our laser. This setup enables us to actively isolate the vibration source, eliminating distortions introduced by environmental noise. In this experiment, the target speaker plays the audio "The law of the school..." (Fig. 10(c)), while the noise source simultaneously plays a song, creating a challenging audio separation scenario. The spectrogram of the microphone recording in Fig. 10(a) reveals that the target audio is heavily masked by background noise, resulting in a poor PESQ score of 1.06.

In contrast, the spectrogram of our recovered speech (Fig. 10(b)) demonstrates significant improvement. Because

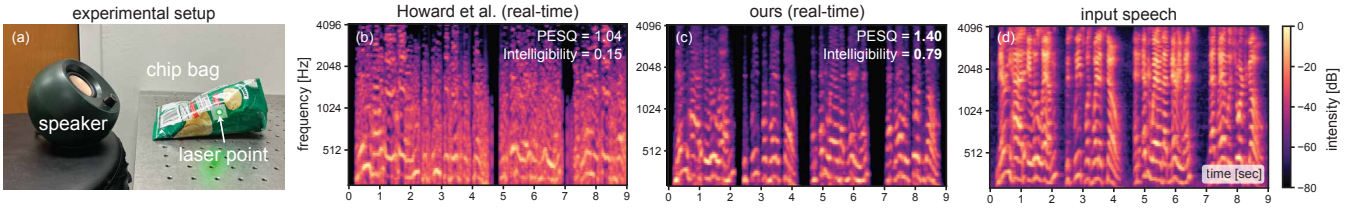

Fig. 8. Comparison of human speech recovery from chip bag with previous methods for real-time reconstruction. **(a)** We replicate the chip bag experiment originally proposed by Davis et al. [8]. In our setup, a speaker plays the "Mary Had a Little Lamb..." audio provided by Davis et al., inducing vibrations in the chip bag. These vibrations then move the speckle pattern, which is captured by our system. **(b)** Recovered audio from Howard et al. [18], which used a zero-crossing algorithm with an event camera for passive sensing. **(c)** Recovered audio using our method. **(d)** Ground truth speech audio.

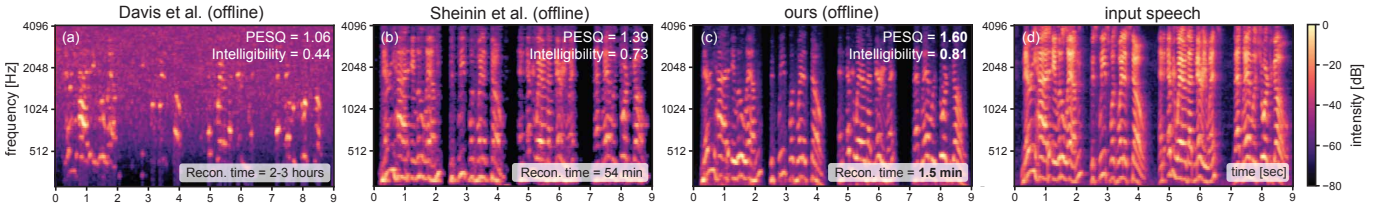

Fig. 9. Comparison of human speech recovery from chip bag with previous offline methods. **(a)** Recovered audio from Davis et al. [8], which employed a high-speed camera along with a strong light source. The reconstruction generally takes 2-3 hours. **(b)** Recovered audio from Sheinin et al. [9], which utilized a cylindrical lens in combination with global shutter and rolling shutter cameras. The reconstruction takes approximately 54 minutes. **(c)** Recovered audio using our method. The reconstruction only takes 1.5 minutes. **(d)** Ground truth speech audio.

vibrometry directly captures vibrations from the target source—remaining unaffected by ambient noise—we are able to extract a much cleaner speech signal. As a result, overall speech clarity is substantially enhanced, achieving a PESQ score of 1.27 and an intelligibility score of 0.80.

5.4.2 Vibrometry Eliminates Echoes

Echoes [62] occur in enclosed spaces where sound waves reflect off surfaces, creating overlapping signals that distort audio clarity. This effect is particularly common in large conference rooms, open offices, and empty halls. As a result, reducing echoes can significantly enhance speech intelligibility, leading to improved communication quality.

Previous approaches [63], [64], [65] for echo elimination primarily relied on post-processing recorded audio using sophisticated echo cancellation algorithms. These methods are often computationally intensive and may struggle to isolate the original signal from complex, overlapping reflections—particularly in highly reverberant environments or when echo characteristics vary dynamically [66]. To overcome these limitations, we propose a vibrometry-based approach that eliminates the effects of echo by directly sensing surface vibrations, thereby bypassing the acoustic interference present in the surrounding environment.

In Fig. 10(middle), we present our experimental setup inside an empty room, where a speaker plays the audio "A laudable regard..." (Fig. 10(f)). As sound waves are emitted, they reflect off the walls, producing echoes in the form of delayed and overlapping versions of the original sound.

To mitigate the impact of echoes, we directed a laser at the target speaker to measure its physical vibrations, enabling the recovery of speech signals to be unaffected by echo-induced distortions. Figure 10(e) demonstrates fine spectral structure closely matching the input speech signal, whereas the microphone recording (Fig. 10(d)) exhibits spectral smearing along the time axis due to delayed reflections of echoes. Additionally, the repeated echoes gradually decrease in intensity, forming a fading trail that further degrades audio quality.

Since it is difficult to acquire an ideal echo chamber, echoes preserved most spectral features and did not severely distort phonetic content in the recording. Therefore, intelligibility (STOI) remains relatively high (0.73) for the echoed recording with our recovery in 0.81. Besides, this metric primarily evaluates short-time intelligibility and spectral similarity, which are less impacted by moderate echoes. In contrast, PESQ, which evaluates long-term perceptual quality, penalizes echo artifacts, resulting in a greater improvement in our method (1.64) compared to microphone recordings (1.29). To further quantify this effect and fully demonstrate the advantage of vibrometry in echo-prone environments, future work will conduct experiments in a controlled echo chamber to refine and validate the system's performance.

5.4.3 Underwater Vibrometry

Detecting underwater sound is crucial for various scientific and industrial applications (e.g. underwater communication [67] and marine biology monitoring [68]). However, achieving high-quality underwater audio recovery remains a significant challenge due to strong signal attenuation, limited detectable bandwidth, and the multi-path propagation of sound waves. Given that vibrations remain preserved underwater, vibrometry appears to be a promising alternative for underwater audio recovery.

Existing underwater vibrometry methods [69], [70] are either costly or require physical contact, leading to challenges in installation complexity and material compatibility. Moreover, effective vibrometry techniques for recovering

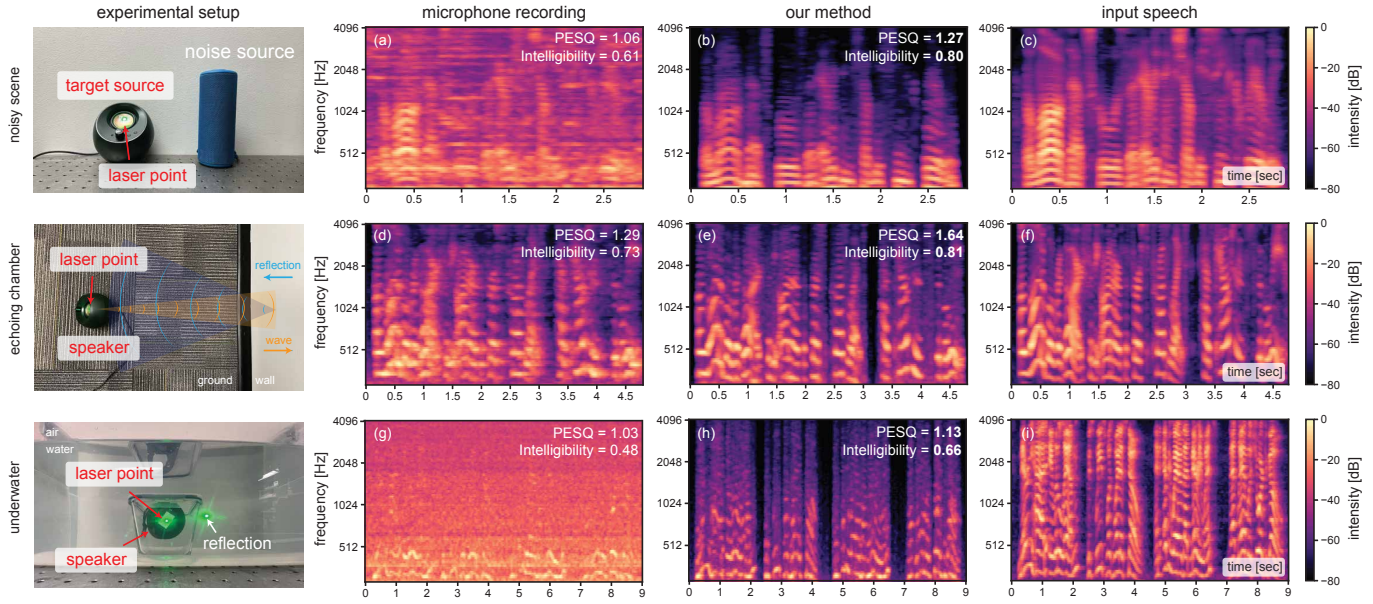

Fig. 10. Vibrometry against environmental distortion. **(a-c)** We construct a noisy environment where the target speaker is influenced by environmental sources. To isolate the target audio signal, we direct a laser onto the audio source playing the speech "The law of the school...", while another speaker plays interfering background noise. The microphone records a mixture of the target speech and noise, whereas the vibrometry measurement successfully isolates the speech from the noisy scene. **(d-f)** We position the system in an empty room to simulate an echo chamber with the speaker playing the audio "A laudable regard...", where reverberation degrades the audio quality. The echo introduces a "delay" effect in the microphone recording, reducing audio resolution in the spectrogram. In contrast, the vibrometry reconstruction preserves the fine details of the spectrogram. **(g-i)** We place a waterproof speaker playing "Marry had a little lamb..." inside a transparent box filled with water, which significantly attenuates the energy of the audio signal, allowing only a small fraction to be captured by the microphone. However, using vibrometry we successfully recovered the speech signal despite the distortion caused by the water.

complex underwater audio signals, such as human speech, have yet to be developed.

We conducted a proof-of-concept experiment to demonstrate the feasibility of underwater audio recovery using our method. As shown in Fig. 10(bottom), we positioned our system on a table and placed a waterproof speaker inside a sealed transparent box filled with water. To prevent it from floating, we secured the speaker inside a smaller water-filled box, anchoring it at the bottom of the enclosure. To capture the vibrations, we directed a laser onto the speaker membrane. While some laser light was reflected by the box surface, creating an additional laser point, this reflection did not align with the speckle pattern originating from the speaker membrane and therefore had a negligible impact on the measurement process.

Although underwater factors such as optical absorption, turbulence, and other unpredictable conditions can influence performance, we assume light reflection loss and water resistance on the speaker membrane to be dominant in this proof-of-concept experiment. More discussion of underwater distortions is provided in the Discussion section.

In Fig. 10(g), the microphone recording captures primarily low-frequency signals, as most high-frequency components are lost after propagating through the water ([PESQ, Intelligibility] = [1.03, 0.48]). In contrast, our method (Fig. 10(h)), which directly measures the physical vibrations of the speaker membrane, achieves improved audio reconstruction ([PESQ, Intelligibility] = [1.13, 0.66]). Although the laser signal also experiences absorption, attenuation, and minor turbulence caused by the water, it provides superior reconstruction quality compared to microphone recordings.

6 Discussion

Light efficiency: The amount of light captured by the sensor significantly impacts the performance of our method. Similar to previous work [9], we used retro-reflective markers to enhance the speckle signal. However, unlike their approach, our method does not require light splitting, effectively doubling the number of photons available for detection. Additionally, while higher-power lasers [24] can further enhance performance, the inherently high dynamic range of event sensors [7] (140 dB vs. 60 dB in standard cameras) reduces the dependence on laser power, offering a more adaptable sensing approach.

Handling large object motion: Subtle vibrations are often superimposed with large-scale object or scene motion when capturing vibrations from musical instruments [9] or hand-held objects. Event cameras are well-suited for capturing both subtle vibrations and large motions simultaneously. Additional experiments that evaluate our system under substantial object motion are provided in the supplementary material.

Event sensitivity: The sensitivity of the event sensor impacts its performance across different levels of motion and can be adjusted through bias settings. In our system, we used the default bias setting for all experiments. Careful tuning of these parameters could enhance the signal-to-noise ratio, improve sensitivity to speckle motion, and ultimately lead to higher-quality audio reconstruction.

Motion dynamic range: Different from traditional cameras, the motion dynamic range of an event camera is influenced by its contrast sensitivity threshold. This dependence can present challenges when attempting to detect subtle

motion below the threshold. In our experiments, system performance degrades as the magnitude of subtle motion decreases. Additional results that quantify performance over a range of motion amplitudes are provided in the supplementary material.

Underwater distortion: We performed a proof-of-concept experiment to recover audio signals underwater. However, distortions in real underwater environments could be far more severe. Factors such as turbulence and light scattering, which were negligible compared to the signal in our experiment, may significantly affect the final measurement. To address these challenges, potential solutions could be adapted from dynamic scattering correction techniques [71], [72], [73].

Event-based flow processing: We explored one possible event flow algorithm for real-time audio recovery. This approach allowed for real-time reconstruction speed while achieving high-quality reconstruction. However, the algorithm remains sensitive to noise and requires hyperparameter tuning. Nevertheless, event-based optical flow tracking algorithms are advancing rapidly [7], [44], [74], and as these methods continue to improve, our event-based vibrometry approach is expected to become more robust and faster to run.

Projecting global 2D motion to 1D: Since different frequencies may oscillate in different directions, there could be improvements to the projection technique by operating in the Fourier domain and having a frequency-adaptive projection method. If we assume that there is one oscillation direction at any given time, it may still be beneficial to use a time-adaptive projection direction (perhaps incremental PCA with a forgetting factor [75]). We found that in practice, projecting on the first (non-time-adaptive) principal component matched the performance of our current method, so we opted for the simpler approach of averaging.

Audio denoising and enhancement: Recovering audio from vibrations inherently introduces noise into the signal [8]. Additionally, materials naturally act as low-pass filters, attenuating high-frequency audio components [4]. Thus, denoising and enhancement post-processing are essential for improving audio quality. However, most learning-based methods [58] are trained on conventional noisy audio datasets, which have different noise characteristics than vibration-induced noise. Future work should focus on training denoising and enhancement models on vibration-derived audio, optimizing both performance and reconstruction quality.

7 Conclusion

We propose a novel high-speed optical vibrometry approach using an event camera. Our system is simple and compact, featuring a fast audio recovery algorithm. We demonstrate its effectiveness across various audio sources, including human speech, in different environments. Our experiments showed additional evidence that optical vibrometry can successfully capture audio in difficult scenarios, like underwater, where traditional microphones suffer. We hope to see future work continue to improve the speed and robustness of vibrometry for live sensing of imperceptible signals.

Acknowledgments

The authors would like to thank Leyla Kabuli, Ethan Weber, Mark Sheinin, Matan Kichler, Rev. zFsZ, Rev. C4Q6, and Rev. uFwf for fruitful discussion, Ruizhi Cao and Tingle Li for manuscript review. This work was supported by STROBE: A National Science Foundation Science & Technology Center under Grant No. DMR 154892 and Weill Neurohub Investigators Program Holographic all-optical electrophysiology: A new platform for ultra-fast bidirectional brain machine interfaces. This material is also based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1752814 (DG). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Additionally, DG was funded by the Center for Innovation in Vision and Optics. The authors thank the developers of the software packages used in this project and not mentioned in the main text, including PyTorch [76], NumPy [77], SciPy [78], matplotlib [79], pesq [50] and librosa [80].

References

[1] C. Cristalli, N. Paone, and R. Rodríguez, "Mechanical fault detection of electric motors by laser vibrometer and accelerometer measurements," *Mechanical Systems and Signal Processing*, vol. 20, no. 6, pp. 1350–1361, 2006.

[2] J. Vass, R. Šmíd, R. Randall *et al.*, "Avoidance of speckle noise in laser vibrometry by the use of kurtosis ratio: Application to mechanical fault diagnostics," *Mechanical Systems and Signal Processing*, vol. 22, no. 3, pp. 647–671, 2008.

[3] N. Wu and S. Haruyama, "Fast motion estimation of one-dimensional laser speckle image and its application on real-time audio signal acquisition," in *Proceedings of the 6th International Conference on Communication and Information Processing*, 2020, pp. 128–134.

[4] K. L. Bouman, B. Xiao, P. Battaglia, and W. T. Freeman, "Estimating the material properties of fabric from video," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1984–1991.

[5] B. T. Feng, A. C. Ogren, C. Daraio, and K. L. Bouman, "Visual vibration tomography: Estimating interior material properties from monocular video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 231–16 240.

[6] M. Schewe and C. Rembe, "Analyzing real-time capability of raw laser-Doppler vibrometer signal combination for signal diversity," in *Optical Measurement Systems for Industrial Inspection XII*, P. Lehmann, W. Osten, and A. A. G. Jr., Eds., vol. 11782, International Society for Optics and Photonics. SPIE, 2021, p. 117820E. [Online]. Available: https://doi.org/10.1117/12.2592048

[7] G. Gallego, T. Delbrück, G. Orchard *et al.*, "Event-based vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 154–180, 2020.

[8] A. Davis, M. Rubinstein, N. Wadhwa *et al.*, "The visual microphone: passive recovery of sound from video," *ACM Trans. Graph.*, vol. 33, no. 4, Jul. 2014. [Online]. Available: https://doi.org/10.1145/2601097.2601119

[9] 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.

[10] P. Lutzmann, B. Göhler, F. Van Putten, and C. Hill, "Laser vibration sensing: overview and applications," *Electro-Optical Remote Sensing, Photonic Technologies, and Applications V*, vol. 8186, pp. 11–26, 2011.

[11] Z. Zalevsky, Y. Beiderman, I. Margalit *et al.*, "Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern," *Optics express*, vol. 17, no. 24, pp. 21 566–21 580, 2009.

[12] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 x 128 120db 30mw asynchronous vision sensor that responds to relative intensity change," in *2006 IEEE International Solid State Circuits Conference-Digest of Technical Papers*. IEEE, 2006, pp. 2060–2069.

[13] Prophesee, "Event-based sensing enables a new generation of machine vision solutions," Prophesee, Tech. Rep., May 2022. [Online]. Available: https://www.prophesee.ai/wp-content/uploads/2022/05/PROPHESEE-White_Paper_Event_Based_Vision_EN_05_09_2022.pdf

[14] R. Guo, Q. Yang, A. S. Chang *et al.*, "Eventlfm: event camera integrated fourier light field microscopy for ultrafast 3d imaging," *Light: Science & Applications*, vol. 13, no. 1, Jun. 2024. [Online]. Available: http://dx.doi.org/10.1038/s41377-024-01502-5

[15] O. Buyukozturk, J. G. Chen, N. Wadhwa *et al.*, "Smaller than the eye can see: Vibration analysis with video cameras," in *19th World Conference on Non-Destructive Testing 2016 (WCNDT)*, vol. 1, 2016.

[16] A. Davis, K. L. Bouman, J. G. Chen *et al.*, "Visual vibrometry: Estimating material properties from small motion in video," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 5335–5343.

[17] J. G. Chen, A. Davis, N. Wadhwa *et al.*, "Video camera–based vibration measurement for civil infrastructure applications," *Journal of Infrastructure Systems*, vol. 23, no. 3, p. B4016013, 2017.

[18] M. Howard and K. Hirakawa, "Event-based visual microphone," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[19] R. Niwa, T. Fushimi, K. Yamamoto, and Y. Ochiai, "Live demonstration: Event-based visual microphone," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4054–4055.

[20] C. Liu, A. Torralba, W. T. Freeman, F. Durand, and E. H. Adelson, "Motion magnification," *ACM transactions on graphics (TOG)*, vol. 24, no. 3, pp. 519–526, 2005.

[21] M. Elgharib, M. Hefeeda, F. Durand, and W. T. Freeman, "Video magnification in presence of large motions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4119–4127.

[22] B. Y. Feng, H. Alzayer, M. Rubinstein, W. T. Freeman, and J.-B. Huang, "3d motion magnification: Visualizing subtle motions from time-varying radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9837–9846.

[23] R. F. Voss and J. Clarke, "'1/fnoise' in music and speech," *Nature*, vol. 258, no. 5533, pp. 317–318, 1975. [Online]. Available: https://doi.org/10.1038/258317a0

[24] S. Bianchi and E. Giacomozzi, "Long-range detection of acoustic vibrations by speckle tracking," *Applied optics*, vol. 58, no. 28, pp. 7805–7809, 2019.

[25] J. W. Goodman, *Speckle phenomena in optics: theory and applications*. Roberts and Company Publishers, 2007.

[26] M. Alterman, C. Bar, I. Gkioulekas, and A. Levin, "Imaging with local speckle intensity correlations: theory and practice," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 3, pp. 1–22, 2021.

[27] K. Jo, M. Gupta, and S. K. Nayar, "Spedo: 6 dof ego-motion sensor using speckle defocus imaging," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4319–4327.

[28] B. M. Smith, P. Desai, V. Agarwal, and M. Gupta, "Colux: Multi-object 3d micro-motion analysis using speckle imaging," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.

[29] J. Zizka, A. Olwal, and R. Raskar, "Specklesense: fast, precise, low-cost and compact motion sensing using laser speckle," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 489–498.

[30] D. Zhu, L. Yang, Z. Li, and H. Zeng, "Remote speech extraction from speckle image by convolutional neural network," in *2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2020, pp. 1–6.

[31] N. Wu and S. Haruyama, "The 20k samples-per-second real time detection of acoustic vibration based on displacement estimation of one-dimensional laser speckle images," *Sensors*, vol. 21, no. 9, p. 2938, 2021.

[32] T. Zhang, M. Sheinin, D. Chan *et al.*, "Analyzing physical impacts using transient surface wave imaging," in *Proc. IEEE CVPR*, 2023.

[33] M. Mahowald and M. Mahowald, "The silicon retina," *An Analog VLSI System for Stereoscopic Vision*, pp. 4–65, 1994.

[34] Y. Chen, S. Guo, F. Yu *et al.*, "Event-based motion magnification," 2024. [Online]. Available: https://arxiv.org/abs/2402.11957

[35] A. N. Angelopoulos, J. N. Martel, A. P. Kohli, J. Conradt, and G. Wetzstein, "Event based, near eye gaze tracking beyond 10,000 hz," *arXiv preprint arXiv:2004.03577*, 2020.

[36] Y. Wang, R. Idoughi, and W. Heidrich, "Stereo event-based particle tracking velocimetry for 3d fluid flow reconstruction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. Springer, 2020, pp. 36–53.

[37] Z. Ge, P. Zhang, Y. Gao, H. K.-H. So, and E. Y. Lam, "Lens-free motion analysis via neuromorphic laser speckle imaging," *Optics Express*, vol. 30, no. 2, pp. 2206–2218, 2022.

[38] B. Oliver, "Sparkling spots and random diffraction," *Proceedings of the IEEE*, vol. 51, no. 1, pp. 220–221, 1963.

[39] J. W. Goodman, "Statistical properties of laser speckle patterns," in *Laser speckle and related phenomena*. Springer, 1975, pp. 9–75.

[40] J. C. Dainty, *Laser speckle and related phenomena*. Springer science & business Media, 2013, vol. 9.

[41] Y. V. Pershin and M. Di Ventra, "Memory effects in complex materials and nanoscale systems," *Advances in Physics*, vol. 60, no. 2, pp. 145–227, 2011.

[42] G. Osnabrugge, R. Horstmeyer, I. N. Papadopoulos, B. Judkewitz, and I. M. Vellekoop, "Generalized optical memory effect," *Optica*, vol. 4, no. 8, pp. 886–892, 2017.

[43] R. Cao, D. Galor, A. Kohli, J. L. Yates, and L. Waller, "Noise2image: noise-enabled static scene recovery for event cameras," *Optica*, vol. 12, no. 1, pp. 46–55, Jan 2025. [Online]. Available: https://opg.optica.org/optica/abstract.cfm?URI=optica-12-1-46

[44] R. Benosman, S.-H. Ieng, C. Clercq, C. Bartolozzi, and M. Srinivasan, "Asynchronous frameless event-based optical flow," *Neural Networks*, vol. 27, pp. 32–37, 2012.

[45] M. Ikura, C. Le Gentil, M. G. Müller *et al.*, "Rate: Real-time asynchronous feature tracking with event cameras," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 11 662–11 669.

[46] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*. Springer, 2003, pp. 363–370.

[47] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[48] T. Sainburg, "timsainb/noisereduce: v1.0," Jun. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3243139

[49] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, p. e1008228, 2020.

[50] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.

[51] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[52] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, pp. 125–128 vol.1.

[53] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993.

[54] R. V. Cox, S. F. De Campos Neto, C. Lamblin, and M. H. Sherif, "Itu-t coders for wideband, superwideband, and fullband speech communication [series editorial]," *IEEE Communications Magazine*, vol. 47, no. 10, pp. 106–109, 2009.

[55] R. Solovyev, A. Stempkovskiy, and T. Habruseva, "Benchmarks and leaderboards for sound demixing tasks," *arXiv preprint arXiv:2305.07489*, 2023.

[56] Y. Mitsufuji, G. Fabbro, S. Uhlich *et al.*, "Music demixing challenge 2021," *Frontiers in Signal Processing*, vol. 1, p. 808395, 2022.

[57] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Transactions on Signal processing*, vol. 56, no. 5, pp. 1830–1839, 2008.

[58] H. Purwins, B. Li, T. Virtanen *et al.*, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[59] J. Xie, J. G. Colonna, and J. Zhang, "Bioacoustic signal denoising: a review," *Artificial Intelligence Review*, vol. 54, pp. 3575–3597, 2021.

[60] X. Liu, Q. Kong, Y. Zhao *et al.*, "Separate anything you describe," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[61] H. Wang, J. Hai, Y.-J. Lu *et al.*, "Soloaudio: Target sound extraction with language-oriented audio diffusion transformer," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[62] D. W. Ricker, *Echo signal processing*. Springer Science & Business Media, 2012, vol. 725.

[63] M. Sondhi, "An adaptive echo canceller," *Bell System technical journal*, vol. 46, no. 3, pp. 497–511, 1967.

[64] A. Mader, H. Puder, and G. U. Schmidt, "Step-size control for acoustic echo cancellation filters–an overview," *Signal Processing*, vol. 80, no. 9, pp. 1697–1719, 2000.

[65] C. Paleologu, J. Benesty, and S. Ciochina, "Sparse adaptive filters for echo cancellation," 2011.

[66] J. Benesty, T. Gänsler, D. R. Morgan *et al.*, "Advances in network and acoustic echo cancellation," 2001.

[67] A. H. Bass and C. W. Clark, "The physical acoustics of underwater sound communication," in *Acoustic communication*. Springer, 2003, pp. 15–64.

[68] A. D. Hawkins, "Underwater sound and fish behaviour," in *The behaviour of teleost fishes*. Springer, 1986, pp. 114–151.

[69] A. R. Harland, J. N. Petzing, and J. R. Tyrer, "Visualising scattering underwater acoustic fields using laser doppler vibrometry," *Journal of sound and vibration*, vol. 305, no. 4-5, pp. 659–671, 2007.

[70] ——, "Nonperturbing measurements of spatially distributed underwater acoustic fields using a scanning laser doppler vibrometer," *The Journal of the Acoustical Society of America*, vol. 115, no. 1, pp. 187–195, 2004.

[71] S. Gigan, O. Katz, H. B. De Aguiar *et al.*, "Roadmap on wavefront shaping and deep imaging in complex media," *Journal of Physics: Photonics*, vol. 4, no. 4, p. 042501, 2022.

[72] H. Ruan, Y. Liu, J. Xu, Y. Huang, and C. Yang, "Fluorescence imaging through dynamic scattering media with speckle-encoded ultrasound-modulated light correlation," *Nature Photonics*, vol. 14, no. 8, pp. 511–516, 2020.

[73] C. Yi, J. Jung, J. Im *et al.*, "Single-shot temporal speckle correlation imaging using rolling shutter image sensors," *Optica*, vol. 9, no. 11, pp. 1227–1237, 2022.

[74] S. Shiba, Y. Aoki, and G. Gallego, "Secrets of event-based optical flow," in *European Conference on Computer Vision*. Springer, 2022, pp. 628–645.

[75] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, pp. 125–141, 05 2008.

[76] A. Paszke, S. Gross, F. Massa *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[77] C. R. Harris, K. J. Millman, S. J. van der Walt *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020.

[78] P. Virtanen, R. Gommers, T. E. Oliphant *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[79] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in science & engineering*, vol. 9, no. 03, pp. 90–95, 2007.

[80] B. McFee, C. Raffel, D. Liang *et al.*, "librosa: Audio and music signal analysis in python," 01 2015, pp. 18–24.

Mingxuan Cai is a Ph.D. student in Electrical Engineering and Computer Sciences at the University of California, Berkeley, advised by Prof. Laura Waller. He received the B.S. degree in Optical Engineering from Zhejiang University, Hangzhou, China, in 2022. His current research interest focuses on space-time imaging, event-based vision, and computational fluorescence imaging.

Dekel Galor is a Ph.D. student in Electrical Engineering and Computer Sciences at the University of California, Berkeley, where he is co-advised by Prof. Laura Waller and Prof. Jacob Yates. He received the B.S. degree in Electrical Engineering and Computer Sciences from UC Berkeley in 2022. His research lies at the intersection of computational imaging, visual neuroscience, and theoretical machine learning.

Amit Pal Singh Kohli is a 5th year Ph.D. student in Laura Waller's group at UC Berkeley. He did his B.S. in Electrical Engineering at Stanford University prior to coming to Berkeley. He works at the intersection of optics, statistics, and machine learning to build robust computational imaging systems for biomedical applications.

Jacob L. Yates is an Assistant Professor in the Herbert Wertheim School of Optometry and Vision Science at UC Berkeley. He is affiliated with the Helen Wills Neuroscience Institute and the Redwood Center for Theoretical Neuroscience. He received his PhD from UT Austin in 2016 and was a postdoctoral researcher at the University of Rochester and the University of Maryland, College Park.

Laura Waller is the Charles A. Desoer Professor of Electrical Engineering and Computer Sciences at UC Berkeley. She received BS, MEng and PhD degrees from the Massachusetts Institute of Technology in 2004, 2005 and 2010. After that, she was a Postdoctoral Researcher and Lecturer of Physics at Princeton University from 2010-2012. She is a Packard Fellow for Science & Engineering, OSA Fellow, and Chan-Zuckerberg Biohub Investigator.

Supplementary Material

S1 Additional Experimental Results

S1.1 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. S1(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. S1(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.

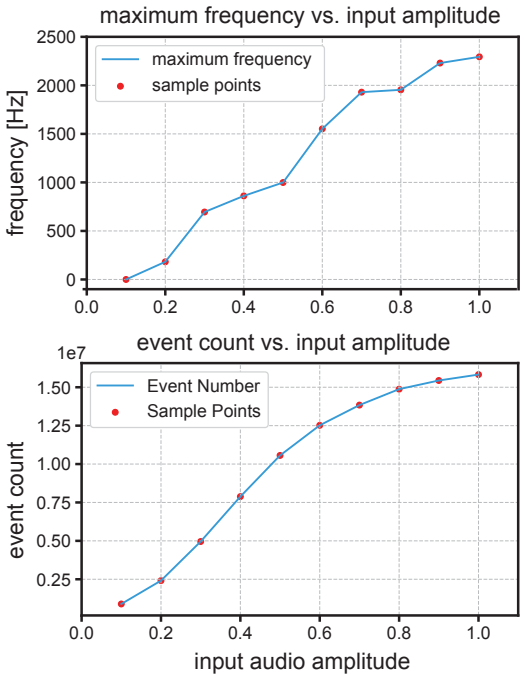

Fig. S1. Experimental results of motion dynamic range evaluation. **(Top)** The maximum reconstructible frequency with different input audio amplitudes from 0.1 to 1.0. **(Bottom)** The recorded event count with different input audio amplitudes from 0.1 to 1.0.

S1.2 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. S2(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. S2(a)), which is comparable to the large motion reported in [9].

When zooming in on the motion trajectory in Fig. S2(a), the waveform corresponding to the 440 Hz signal becomes clearly visible. Consequently, in the reconstruction shown in Fig. S2(c), we successfully recovered the 440 Hz audio despite the presence of substantial scene motion.

S1.3 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. S3(b–d)) while simultaneously recording the mixed audio with a microphone. As shown in Fig. S3(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. S3(e–g).

Other than audio source separation, Zhang *et al.* [32] demonstrated that having more laser spots enables the analysis of surface vibration waves. This in turned allowed for impact localization and analysis of material properties, which are exciting avenues for future exploration of event-based vibrometry.

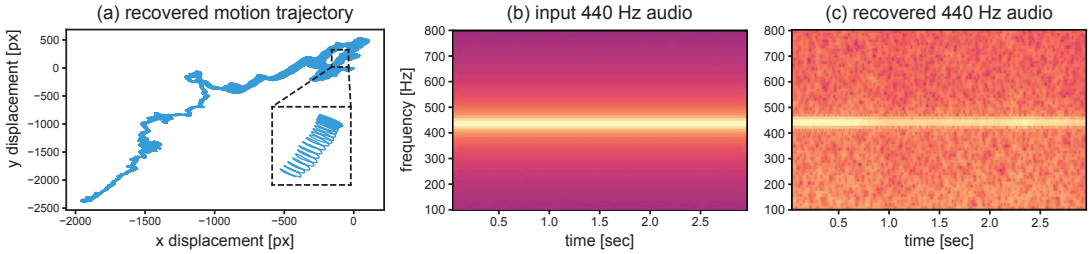

Fig. S2. Experiment results with large hand-holding motion. **(a)** Recovered motion trajectory. **(b)** Input 440 Hz audio. **(c)** Recovered 440 Hz audio.

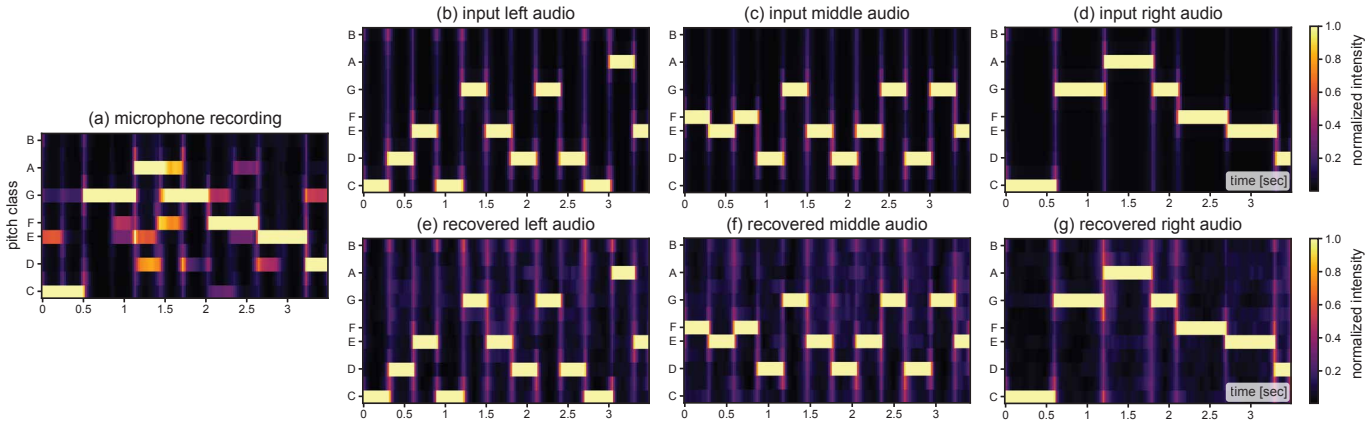

Fig. S3. Capturing signals from three laser spots and audio sources. **(a)** The chromagram of the microphone recording. **(b-d)** The chromagrams of the input audio from left, middle, and right speakers. **(e-g)** The chromagrams of the recovered left, middle, and right audio signals.