Event-based Visual Microphone using Laser Speckle

Ryo Shirakawa[1], Shiori Sugimoto[1], Yoko Sogabe[1], Shoichiro Saito[1], and Masaki Kitahara[1]

[1]*NTT, Inc.* , Kanagawa, Japan
[1]{ryo.shirakawa, shiori.sugimoto, yoko.sogabe, shoichiro.saito, masaki.kitahara}@ntt.com

***Abstract*—Since sound causes minute vibrations in objects surrounding the sound source, Visual Vibrometer allows us to recover the sound from remote point by optically measuring the vibration on the object's surface. For that system, Laser Doppler Velocimeters (LDVs) and high-speed cameras are commonly used, but there are issues in terms of equipment cost and data efficiency. Therefore, a technique has recently emerged to use event-based cameras in place of such equipment. Event-based cameras record only changes in brightness, independently, at each pixel, and its advantages such as high temporal resolution, high data efficiency, and simple device structure make it easier to measure. However, the technique is highly dependent on the object's surface characteristics, so the measurement conditions can prove difficult to realize. In this paper, we propose a new measurement system that observes changes in laser speckles caused by vibration using an event-based camera to achieve more robust measurement conditions, and a method for recovering sound from the event signals. The proposed sound recovery algorithm recovers the audio signal by noise reduction, sign assignment, and integration, assuming that the number of events produced by speckle pattern shift and detected at each time is closely related to the absolute value of the vibration speed. This method is very simple yet effective, and its performance is demonstrated by experiments in real environments.**

***Index Terms*—remote sound acquisition, visual acoustics, event-based camera, laser speckle**

I. Introduction

Sound is a fluctuation of pressure in the atmosphere, and when it hits an object, it exerts a force that causes the object's surface to deform as determined by its own vibration mode. These vibrations are usually invisible to the human eye because the spatial fluctuations are so small, but they contain enough information that the sound can be recovered.

A Visual Vibrometer [1]–[8] is a system that uses optical equipment to measure the minute vibrations present on the surface of an object; it is used for measuring the physical properties of materials [4], detecting abnormalities in equipment [6], and so on. Using this system, it is possible to acquire vibration on the surface of an object from a remote point and recover the sound at that location. Sound recovery using visual information is superior to conventional microphones in terms of attenuation and directivity, and the expectation is that it will be used in important applications such as surveillance and security. These require measurement devices with high spatial and temporal resolution, and for this reason, Laser Doppler Velocimeters (LDVs) and high-speed cameras are used, but issues remain regarding the cost of the devices and how to process the large amounts of data acquired.

Event-based cameras [9] are unique imaging devices designed to mimic the retina of living organisms; they record only changes in brightness, independently, at each pixel. Since these sensors output data only from pixels that experience changes in brightness, even when imaging at extremely high spatial and temporal resolutions, the data stream will be very sparse. Frame cameras suffer from a paucity of light per pixel if the spatial resolution is increased, which increases the exposure time to compensate, so the temporal resolution will decrease. With event cameras, however, there is no such trade-off, because even if the quantity of light per pixel is low, the detection sensitivity can be sufficiently lowered to maintain event detection accuracy. As a result, it is possible to acquire high-frequency data that satisfies the Nyquist rate of audible sound with a higher spatial resolution than is possible with high-speed cameras. In addition, because device structure is simple and compact, it is possible to perform measurements more simply than allowed by conventional devices.

However, if the sound level is low or the frequency is high, the vibration of the object's surface becomes so minute that even event cameras may fail to capture the surface vibration. In such cases, laser speckle contrast imaging (LSCI) [10]–[12] is useful. Speckle is a phenomenon in which a noise-like pattern is produced by the interference of scattered light when a rough surface is irradiated by coherent light. Since speckle patterns are very sensitive to deformation of the irradiated surface, they support high-resolution measurements in the medical [10] and industrial [12] fields.

In this paper, we propose a new measurement system that observes changes in laser speckle caused by vibration using an event-based camera to achieve more robust measurements, and a method for recovering sound from the event signals. The proposed sound recovery algorithm recovers the audio signal by noise reduction, sign assignment, and integration; its assumption is that the number of events produced by speckle pattern shift and detected at each time is closely related to the absolute value of the vibration speed. Our proposal allows high-frequency sound to be captured with inexpensive equipment, and it achieves sufficient performance for practical use. Finally, we demonstrate the performance of the proposed system through experiments in a real environment using a bag of chips and a speaker.

A. Related work

In recent years, there has been extensive research into using imaging devices for remote sound recovery. The Visual

Microphone (VM) [3], a representative work in this field, uses a high-speed camera to recover sound in a completely passive manner. With this method, the performance of sound recovery with high-speed camera depends on whether its temporal resolution meets the required sampling rate for the sound frequency and if its spatial resolution is high enough to capture the vibration's magnitude. Therefore, this method demands a camera with very high performance, and a carefully selected lens; its measurement conditions are rather restrictive. In addition, for higher performance, it is necessary to process a large amount of data, making the computation time impractically long. A simple alternative method has been proposed that uses a conventional video camera with a CMOS sensor and a rolling shutter. This method has improved temporal resolution as its rolling shutter captures data at slightly different times from each line of the sensor. However, it is necessary to obtain a reference image of the object in a stable state in advance, or to precisely align the direction of the edge of the object with the direction of the sensor row.

Sheinin et al. [7] designed a new imaging system that combines a rolling shutter camera with a global shutter camera and a beamsplitter; this combination allows the simultaneous measurement of the reference image of the object as well as the surface fluctuations. In addition, as laser speckle allows the capture of even minute vibrations at low spatial resolutions, the restrictions on lenses are greatly relaxed. Although the measurement target is limited to the laser's irradiation point, multiple lasers and the division of the image sensor by using a cylindrical lens make it possible to perform multi-point measurements. The complexity of the imaging system is a limitation of this method.

The Event-Based Visual Microphone (EBVM) [8] uses an event-based camera to map the event signals from the specular reflection from an object to the zero-crossings of the sound signal and recover the sound by projection onto convex sets (POCS) with the requirements being zero-crossing and Fourier support. This method can detect minute vibrations by detecting changes in specular reflection even at low spatial resolution, but its feasibility strongly depends on the surface conditions. Moreover, although the sound recovery algorithm based on zero-crossing can estimate the relative intensity of each frequency from short-time signals, it cannot estimate the absolute intensity, so the *loudness* of the entire signal is not guaranteed.

In this paper, we aim to achieve a simple device structure, object-independent performance, and low computation cost by using an event-based camera and laser speckle.

II. Background

A. Event-based Camera

Event-based cameras detect changes in brightness at each pixel and output them as event signals. Each event is represented as follows.

$$e = (\mathbf{x}, p, t) \tag{1}$$

where $\mathbf{x} = [x_x, x_y] \in \mathbb{Z}^2$ are the coordinates of the pixel at which the event occurred, $p \in \{-1, 1\}$ is the logarithmic brightness gradient, called polarity, and t is the timestamp.

For convenience, we denote a set of events E within a given spatial area X using the following formula.

$$E = \{e(\mathbf{x}_1, p_1, t_1), e(\mathbf{x}_2, p_2, t_2), ..., e(\mathbf{x}_n, p_n, t_n)) \mid \mathbf{x} \in X\} \tag{2}$$

The temporal resolution of event cameras is on the scale of microseconds, which satisfies the Nyquist rate of human audible range. Since the pixel values are output independently, the data efficiency is much higher and the power consumption is lower than this possible with a high-speed camera that outputs all pixels frame-by-frame. The device structure of event-based cameras is similar to that of conventional cameras, except for the image sensor, which is small and lightweight.

B. Speckle Pattern Shift

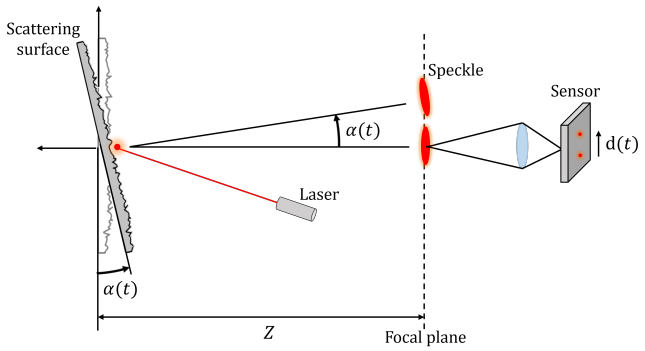


Fig. 1: **Speckle pattern shift.** When a surface is illuminated by coherent light from a laser, light interference generates speckle. When imaging with a strongly defocused camera (the focal plane is positioned far from the surface), changes in the speckle caused by surface deformations can be observed as shifts on the sensor plane.

When coherent light from a laser is irradiated onto an object with a rough surface, it causes scattering within a small irradiated spot. These lights scattered on the surface and reach the sensor with difference phases due to the path differences; they interfere with each other to form a random spatial pattern called speckle pattern. Deformation of the scattering surface within the spot causes a random change in the phases of the scattered lights, resulting in a significant change in the speckle pattern. Due to this characteristic, laser speckle is often used to measure minute changes in objects.

Although the speckle pattern changes depending on the scattering surface and camera conditions [13], it has been shown that if the camera's focal plane is far from the surface, i.e., strongly defocused, the speckle pattern is not affected by changes in the position of the scattering surface, only by changes in the tilt (Fig.1) [14]. Under this condition, the speckle pattern simply shifts on the sensor plane depending on the change in the tilt of the scattering surface. The distance

of the speckle pattern shift $d(t)$ is calculated by the following equation.

$$d(t) = \frac{Z \tan \alpha(t)}{M} \simeq \frac{Z\alpha(t)}{M} \tag{3}$$

where Z is the distance from the scattering surface to the focal plane, $\alpha(t)$ is the tilt of the scattering surface, and M is the inverse of the magnification of the imaging system. There is a linear relationship between the amount of shift of the speckle pattern and the tilt of the scattering surface. When the target vibrates, the speckle pattern also vibrates linearly on the sensor plane. This implies that it may be possible to recover information such as the amplitude and frequency of the vibration by analyzing the shift of the speckle pattern.

III. Methods

As shown in Sec. II-B, the vibration information can be recovered by tracking the speckle pattern captured by the sensor. For this, feature point tracking by optical flow is often used. Although computing methods for optical flow in event signals have been actively researched in recent years [15], [16], it is difficult to use them because the amount of movement is spatially minute and temporally large compared to the moving scenes assumed for in-vehicle cameras and drones.

Therefore, we propose a vibration estimation algorithm based on an analytical approach using the number of event occurrences. For one frame, the total number of events that occur when a pattern moves at a certain speed on a sensor is closely related to the pattern's speed. In this paper, the number of events is treated as the unsigned speed. Let $N(t')$ be the number of events, it can be calculated by following formula.

$$\begin{aligned} N(t') = N(k\Delta t) &= \sum_{t_i \in E} \mathbb{I}(t_i, k) \\ \mathbb{I}(t_i, k) &= \begin{cases} 1 \text{ if } k\Delta t \leq t_i < (k+1)\Delta t \\ 0 \text{ otherwise} \end{cases} \end{aligned} \tag{4}$$

where Δt is the count interval and k is a sampling index, so $t' = k\Delta t$ is a sampling time.

If $v(t)$ is the velocity of speckle pattern shift, the following equation holds.

$$\frac{d}{dt} d(t) = v(t) \tag{5}$$

Since the number of events is related to the pattern's speed, the following relationship holds using Eq. (4) and (5).

$$|v(t')| \propto N(t') \tag{6}$$

$$d(t) \propto \int sign(v(t'))N(t') \tag{7}$$

where $sign(\cdot)$ is the sign function.

From Eq. (7), if the sign of $v(t')$ is determined, the vibration information can be estimated. The sign of velocity cannot be estimated from unsigned speed, but if we assume that the motion is generated by simple oscillation, we can consider that the sign of velocity will reverse at the point where the speed becomes zero. Because of factors such as the sampling rate, it is not always possible to record the moment when the speed becomes zero, so we estimate the timing at which the direction of motion will reverse by relaxation to the local minimum point.

This is a simple idea, but our preliminary experiment showed it can be successfully applied to realistic data. For verification, we simulated signed and unsigned time series data of the vibration speed for three audio sources: Chirp (200-2,000Hz), MIDI ("Mary had a little lamb" from [3]), and Speech ("Mary had a little lamb ..." from [3]). Then, we looked at all the points where the unsigned signal took its local minimum value, and all the points where the sign of the signed signal inverted. Finally, we evaluated the precision of detecting the inversion points as follows:

$$precision = \frac{Set(\text{sign inversion}) \cap Set(\text{local minimum})}{Set(\text{local minimum})} \tag{8}$$

where $Set(\cdot)$ is the set of target points. Here, the local minimum was detected using the scipy argrelmin function. The results were 100% for Chirp, 96.6% for MIDI, and 79.0% for Speech. The more complex the audio signal, the worse the estimation performance, but despite being a very simple method, it is still effective. Also, considering the nature of sound signals, even if the sign assignment is wrong, if the error is not concentrated, it may not affect the overall sound impression because it appears as short time noise.

In real environments, it is difficult to detect local minimum points because the recorded event data contains noise. Therefore, our algorithm uses a hard threshold process based on Fourier transforms to remove just the noise from the target signal and so detect the local minima. The threshold process is achieved by the following policy.

$$\mathcal{N}_l'(\omega) = \begin{cases} \mathcal{N}_l(\omega) \text{ if } |\mathcal{N}_l(\omega)| > \max_{\omega} |\mathcal{N}_l(\omega)| * \eta \\ 0 \text{ otherwise} \end{cases} \tag{9}$$

where $\mathcal{N}_l : \mathbb{R} \to \mathbb{C}$ is the short-time Fourier transform of signal $N_l(t')$ and η is the threshold parameter. Our algorithm recovers the velocity of the speckle pattern shift by performing local minimum detection on the denoised signal and assigning sign inversion to each detected point. Finally, integrating the recovered velocity yields an estimate of the vibration signal. Offset, which occurs due to the imperfect assignment of signs, is corrected by taking the difference against the moving average. Our framework is summarized in Fig. 2.

IV. Experiments

A. Setup

We constructed a measurement system using a Prophesee EVK4 (Sony IMX636ES, 1280 × 720 pixels) with a 50mm lens and a 638nm 40mW laser. The measurement conditions were the same as described in related methods [3], [7], [8], with the bag of chips placed in front of the speaker and captured from 1m away. The three audio sources used in Sec. III were output from the speaker. We conducted experiments on each audio source under two audio level conditions, High and Low, to verify the robustness of the methods tested against

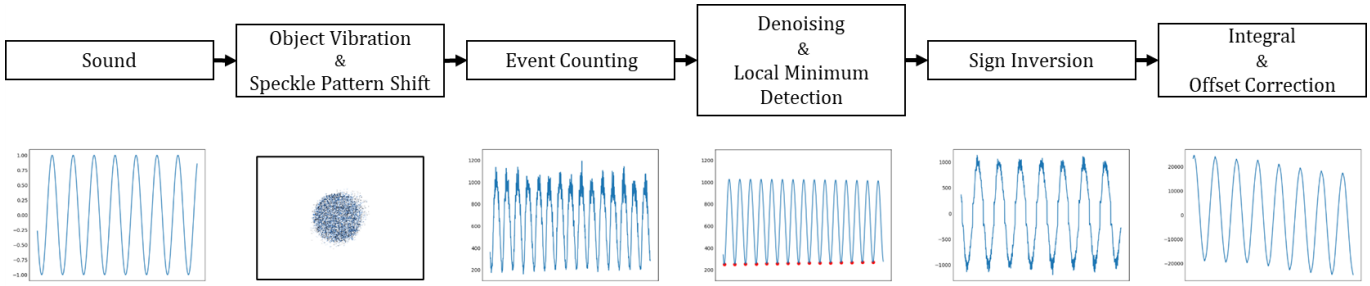


Fig. 2: **Framework overview.** The object vibration caused by sound is observed as speckle pattern shifts by the event camera. Event signals are counted during the count interval and converted into unsigned speed values. The velocity is recovered by sign restoration through denoising and local minimum detection. Finally, by integrating the velocity and offset correction, the vibration signal is estimated.

audio levels. The audio levels were set as follows: average of the entire signal was 90dB(A) and 120dB(A) for Chirp, 75dB(A) and 105dB(A) for MIDI, and 85dB(A) and 115dB(A) for Speech. For our method, parameter Δt was set to $\frac{1}{44{,}100}$ [s] and η to 0.05 for all audio sources.

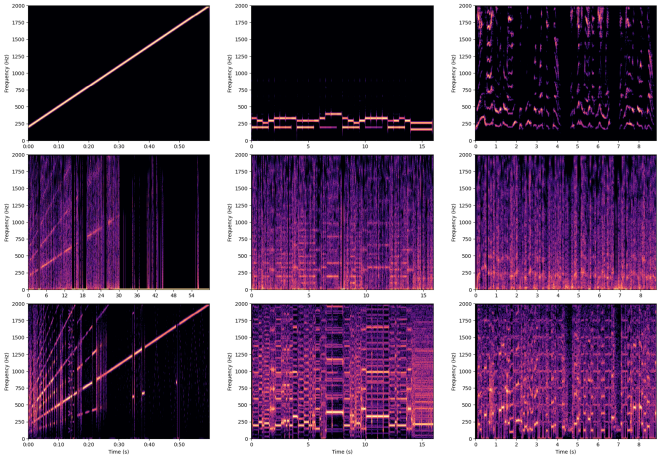
EBVM [8] was used as a benchmark for evaluating the performance of the recovered audio. To capture specular reflection for EBVM, we set a DC powered light to the side of the target object. The Fourier support was set at 5%, and the convergence condition was set to 0.1% or less of the signal for the part that was changed due to the zero crossing constraint.

Performance of each method was evaluated using three metrics: segmental SNR (SSNR) [17], median vector angle error (MVA) [8], and vector angle error for the entire signal (VA). Since the recovered audio signal does not contain the intensity information of the original signal, the signal used for SSNR is normalized using the maximum value of the entire signal.

B. Sound Recovery from a Bag of Chips

Fig. 3 shows the spectrograms of the audio recovered using each method under the High audio level condition. The results of the EVBM recovery were weak in signal strength for all audio sources, and there was a lot of noise overall. Chirp could not be recovered after 1000Hz because the vibration was too minute to observe changes in specular reflection. On the other hand, our method exhibited relatively little noise and could successfully recover the original signal for Chirp and MIDI. Comparing the results of our method for each sound source, it appears that the recovery performance for Speech is low. The performance of the sign inversion estimation described in Sec. III is thought to be one of the causes. Actually, the limitation of our method is that it cannot correctly assign signs if the signal has a stationary point that is neither a local maximum nor a local minimum. For signals with complex waveforms, such as speech, there are many such points, and this is thought to be one factor reducing the accuracy of sign inversion.

Fig. 4 shows the spectrograms of the audio recovered using each method under Low audio level condition. As shown in the middle row of Fig. 4, EVBM fails completely at the

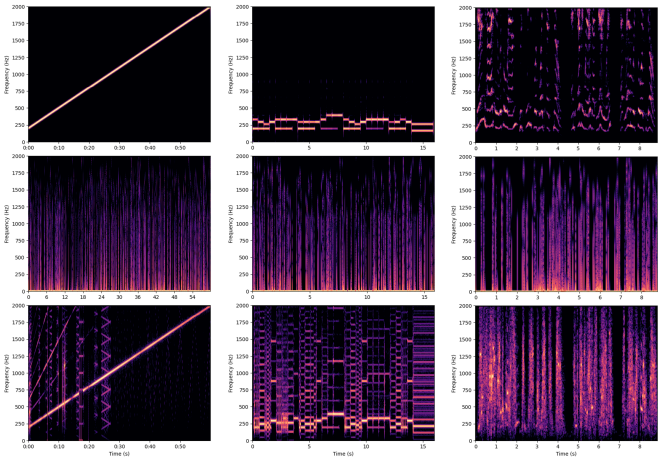


(a) Chirp (b) MIDI (c) Speech

Fig. 3: **Sound recovery under "High" audio level condition.** Spectrograms of (top row) ground truth, (middle row) recovered audio signals by EBVM, (bottom row) recovered audio signals by our method. The audio played was (a) Chirp, (b) MIDI, or (c) Speech.

low audio level. The reason for this is that when the audio level decreases, the vibration of the object also becomes more minute, making it more difficult to observe changes in specular reflection. On the other hand, since speckle is sensitive to minute fluctuations, our method is very robust with respect to audio levels. Comparing the bottom rows of Fig. 3 with those of Fig. 4, the superiority of our method is more significant at low audio level rather than at high audio level. This is thought to be due to the fact that the vibration mode of the object becomes corrupted when the audio level is too high.

Tab. I shows that the recovery performance of our method is superior to EBVM in most metrics. For Speech, EBVM was superior in MVA at the high audio level, but our method was comparable. As with the spectrogram results, comparing the performance for each sound source, the results were poor for Speech. Tab. I shows that that our method has excellent recovery performance even at low audio level, and that it has



(a) Chirp (b) MIDI (c) Speech

Fig. 4: **Sound recovery under "Low" audio level condition.** Spectrograms of (top row) ground truth, (middle row) recovered audio signals by EBVM, (bottom row) recovered audio signals by our method. The audio played was (a) Chirp, (b) MIDI, or (c) Speech.

Table I: Sound recovery quality evaluated using segmental signal-to-noise-ratio (SSNR) [17], median vector angle error (MVA) [8], and vector angle error for the entire signal (VA).

Method	Audio	Audio Level	SSNR	MVA	VA
EBVM [8]	Chirp	High	-0.50	89.9	86.8
Ours	Chirp	High	**-0.17**	**15.5**	**51.6**
EBVM [8]	Chirp	Low	-0.88	89.9	89.6
Ours	Chirp	Low	**-0.04**	**15.8**	**33.6**
EBVM [8]	MIDI	High	**-1.41**	67.7	74.1
Ours	MIDI	High	-2.14	**65.2**	**64.9**
EBVM [8]	MIDI	Low	-3.11	87.9	87.3
Ours	MIDI	Low	**-0.34**	**48.6**	**59.6**
EBVM [8]	Speech	High	-7.64	**68.2**	78.9
Ours	Speech	High	**-6.29**	69.1	**75.7**
EBVM [8]	Speech	Low	-15.70	87.2	88.3
Ours	Speech	Low	**-3.13**	**68.3**	**75.2**

high robustness.

V. Conclusion

The proposed method, which uses an event-based camera to observe the changes in laser speckle caused by surface vibration, has a simpler device structure than conventional methods and achieves robust sound recovery with regard to measurement conditions. Our method has no special requirements such as specular reflection, is robust to sound level, and has a very simple recovery process, so its feasibility is extremely high. The proposed method did not perform well for more complex sounds such as speech, but there is room for improvement by developing more advanced algorithms for sign estimation, a future task.

References

[1] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM transactions on graphics (TOG)*, Vol. 31, No. 4, pp. 1–8, 2012.

[2] H-E Albrecht, Nils Damaschke, Michael Borys, and Cameron Tropea. *Laser Doppler and phase Doppler measurement techniques*. Springer Science & Business Media, 2013.

[3] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham Mysore, Fredo Durand, and William T. Freeman. The Visual Microphone: Passive Recovery of Sound from Video. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, Vol. 33, No. 4, pp. 79:1–79:10, 2014.

[4] Abe Davis, Katherine L Bouman, Justin G Chen, Michael Rubinstein, Fredo Durand, and William T Freeman. Visual vibrometry: Estimating material properties from small motion in video. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 5335–5343, 2015.

[5] S Jj Rothberg, MS Allen, Paolo Castellini, Dario Di Maio, JJJ Dirckx, DJ Ewins, Ben J Halkon, P Muyshondt, Nicola Paone, T Ryan, et al. An international review of laser Doppler vibrometry: Making light work of vibration measurement. *Optics and Lasers in Engineering*, Vol. 99, pp. 11–22, 2017.

[6] Mohamad Hazwan Mohd Ghazali and Wan Rahiman. Vibration Analysis for Machine Monitoring and Diagnosis: A Systematic Review. *Shock and Vibration*, Vol. 2021, No. 1, p. 9469318, 2021.

[7] Mark Sheinin, Dorian Chan, Matthew O'Toole, and Srinivasa G. Narasimhan. Dual-Shutter Optical Vibration Sensing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2023.

[8] Matthew Howard and Keigo Hirakawa. Event-Based Visual Microphone. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[9] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-Based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44, No. 1, pp. 154–180, 2022.

[10] David A. Boas and Andrew K. Dunn. Laser speckle contrast imaging in biomedical optics. *Journal of Biomedical Optics*, Vol. 15, No. 1, p. 011109, 2010.

[11] David Briers, Donald D Duncan, Evan Hirst, Sean J Kirkpatrick, Marcus Larsson, Wiendelt Steenbergen, Tomas Stromberg, and Oliver B Thompson. Laser speckle contrast imaging: theoretical and practical limitations. *Journal of biomedical optics*, Vol. 18, No. 6, pp. 066018–066018, 2013.

[12] Wei An. *Industrial applications of speckle techniques*. Industriell produktion, Stockholm, Sweden, 2002.

[13] Kensei Jo, Mohit Gupta, and Shree K. Nayar. SpeDo: 6 DOF Ego-Motion Sensor Using Speckle Defocus Imaging. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4319–4327, 2015.

[14] Zeev Zalevsky, Yevgeny Beiderman, Israel Margalit, Shimshon Gingold, Mina Teicher, Vicente Mico, and Javier Garcia. Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern. *Opt. Express*, Vol. 17, No. 24, pp. 21566–21580, Nov 2009.

[15] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3867–3876, 2018.

[16] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018.

[17] John H. L. Hansen and Bryan L. Pellom. An effective quality evaluation protocol for speech enhancement algorithms. *5th International Conference on Spoken Language Processing (ICSLP 1998)*, 1998.