PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Noise-reducing sound capture based on exposure-time of still camera

Hiroki SHINDO[1]; Koichi TERANO[2]; Kenta IWAI[3];
Takahiro FUKUMORI[4]; Takanobu NISHIURA[5]

[1,2]Graduate School of Information Science and Engineering, Ritsumeikan University, Japan

[3,4,5]College of Information Science and Engineering, Ritsumeikan University, Japan

ABSTRACT

A visual microphone has been proposed to capture the distant sound. This microphone captures the sound from high frame-rate video by using the pixel difference. It is expected to be applied in surveillance cameras because it is robust to be affected by the sound distance. However, this capturing method is unpractical due to expensive-equipment requirement. In this paper, we attempt to realize inexpensive visual microphones by using a still camera with a CMOS image sensor. A CMOS image sensor shoots the image including rolling-shutter distortion. An image shot with the CMOS sensor has short time displacement information because the sensor sequentially writes image to each line. Therefore, it is possible to capture sound from an image without the high frame-rate video in case of acoustic signals can be extracted from the short time displacement information included in the image. However, lower frequency noise is included in the captured sound using the sensor due to the inclination and distortion of the object to be photographed. This noise depends on the exposure-time of the sensor. We thus propose a noise reducing method for the captured sound by a digital filter considering the exposure-time of the CMOS image sensor. Experimental results show that the proposed method can reduce the noise compared with the original captured sound.

Keywords: Sound capturing, Noise reduction, CMOS image sensor, Still camera

1. INTRODUCTION

Capturing a distant sound is very important on security and rescue. Various microphones such as parabolic microphone and shotgun microphone have been developed for capturing a distant sound by forming the directivity [1]. However, the sound is attenuated in distance and it is difficult to capture the distant sound with conventional microphones. To solve this problem, a visual microphone has been proposed that captures the sound from video [2,3,4]. This microphone can directly measure the vibration of the object caused by distant sound using the pixel difference and extract the sound around the object to be photographed. In this method, the distant attenuation of sound wave is less affected than the conventional microphones. However, this microphone needs high frame-rate video to extract the sound and is an unpractical capturing method. In this paper, we attempt to realize inexpensive visual microphones by using still camera. Hence, we focused on the complementary metal-oxide-semiconductor (CMOS) image sensor with rolling-shutter distortion [5]. The CMOS image sensor is exposed and sequentially writes an image from top to the bottom row of the image. Thus, an image shot with the CMOS image sensor has time information in each line of the element. Therefore, it is possible to capture sound without the high frame-rate video in case of acoustic signals can be extracted from the time information included in an image. However, lower frequency noise is mixed in the captured sound using the CMOS image sensor due to the inclination and distortion of the object to be photographed. This noise depends on the exposure-time of the sensor. Therefore, we propose a method of reducing the noise by designing a digital filter based on exposure-time of the sensor.

[1] is0261xf@ed.ritsumei.ac.jp
[2] is0267rs@ed.ritsumei.ac.jp
[3] iwai18sp@fc.ritsumei.ac.jp
[4] fukumori@fc.ritsumei.ac.jp
[5] nishiura@is.ritsumei.ac.jp

2. SOUND CAPTURE FROM AN IMAGE WITH CMOS IMAGE SENSOR

2.1 Method of Sound Capture from An Image Shot by CMOS Image Sensor

In this paper, we focus on the sound capture method by extracting the sound from an image including rolling-shutter distortion with a CMOS image sensor as an imaging device [6]. The rolling-shutter distortion occurs in each line when a moving object is photographed by using a CMOS image sensor. Figure 1 shows the overview of the sound extraction from an image. Sound propagates as air vibrations. Then, the sound emitted from the sound source vibrates an object near the sound source. At this time, the image can be taken by the still camera with CMOS sensor and includes rolling-shutter distortion. This distortion occurs due to a time difference of the image in each line. Therefore, it is possible to capture the vibration of the object to be photographed by extracting coordinate deviations sequentially from the top to the bottom of the image. The captured vibration matches the vibration of the propagated sound, assuming that there is no attenuation of sound at the surface of the object.

For extracting vibrations from the image, we extract the edge of the image. To extract the edge, the edge of the image is enhanced using a Sobel filter as preprocessing [6]. After detection the edge of the image by the Sobel filter, the image is binarized. The 3×3 neighborhood input image $\boldsymbol{V}(x,y)$ and each directional Sobel kernel $\mathbf{G}_\mathrm{x}$, $\mathbf{G}_\mathrm{y}$ are respectively shown in Eqs. (1) and (2).

$$\boldsymbol{V}(x,y) = \begin{bmatrix} v(x-1,y-1) & v(x,y-1) & v(x+1,y-1) \\ v(x-1,y) & v(x,y) & v(x+1,y) \\ v(x-1,y+1) & v(x,y+1) & v(x+1,y+1) \end{bmatrix}, \tag{1}$$

$$\mathbf{G}_\mathrm{x} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \ \mathbf{G}_\mathrm{y} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \tag{2}$$

where, $v(x,y)$ is amplitude at the position (x,y). The output of Sobel filter is $\boldsymbol{v}'(x,y)$ obtained by Eq. (3).

$$\boldsymbol{v}'(x,y) = \sqrt{(\boldsymbol{V}(x,y) * \mathbf{G}_\mathrm{x})^2 + \left(\boldsymbol{V}(x,y) * \mathbf{G}_\mathrm{y}\right)^2}, \tag{3}$$

where, symbol '$*$' is a convolution operator. The binarized amplitude $\boldsymbol{v}''(x,y)$ at $(x,\ y)$ is obtained by Eq. (4).

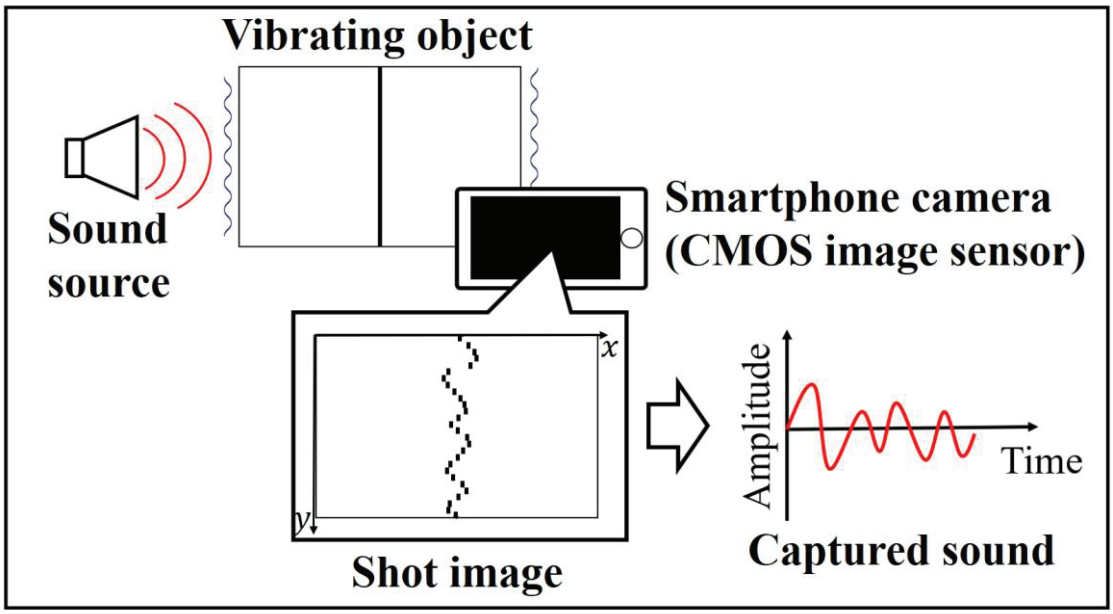


Figure 1 – Overview of the sound capture from an image.

$$\boldsymbol{v}''(x,y) = \begin{cases} (255,255,255) & \text{if } \boldsymbol{v}'(x,y) > \text{ Threshold} \\ (0,0,0) & \text{otherwise} \end{cases}, \tag{4}$$

In this paper, the threshold is determined empirically. The amplitude $s(n)$ of the extracted sound is given by Eq. (5) by using the binary image.

$$s(n) = a(n) - \bar{a}, \quad \left(0, \dots, n, \dots, \frac{F_{out}}{F_{\text{CMOS}}} \times \text{Height}\right), \tag{5}$$

$$a(n) = \arg\min_{x \in \mathbf{C}}(x), \qquad \mathbf{C}\text{: } \boldsymbol{v}''\left(x, \frac{F_{\text{CMOS}}}{F_{\text{out}}} \times n\right) = (255,255,255), \tag{6}$$

where, $a(n)$ is the value of x coordinate of the edge for each line, $\bar{a}$ is average of $a(n)$, n is the sample index, Height is the number of pixels in the y direction of the input image, F_{out} is the sampling frequency of the captured sound and F_{CMOS} is that of CMOS image sensor. Using these equations, we can extract the sound by an image with a CMOS image sensor.

2.2 Results of Sound Capture and Its Problem

We conducted the experiment to confirm that the sound could be extracted from an image. The sound was extracted by using an image which took a paper near the sound source with a CMOS image sensor. We used 0.4 kHz pure tone as sound source emitted from a loudspeaker. Figure 2 shows an input image taken by a CMOS image sensor. Figure 3 shows results of sound capture. Figure 3 (a) and (b) respectively show a waveform and power spectrum of a captured sound. As shown in Fig. 3, we can confirm that a sound with 0.4 kHz can be captured. However, lower frequency noise is mixed into the captured sound. It can be seen from Fig. 3 (b) that the noise power is larger than that of the target sound at 0.4 kHz. Therefore, the noise reduction for the captured sound is required.

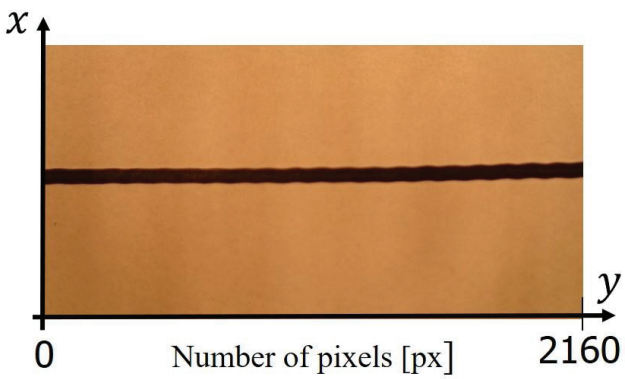
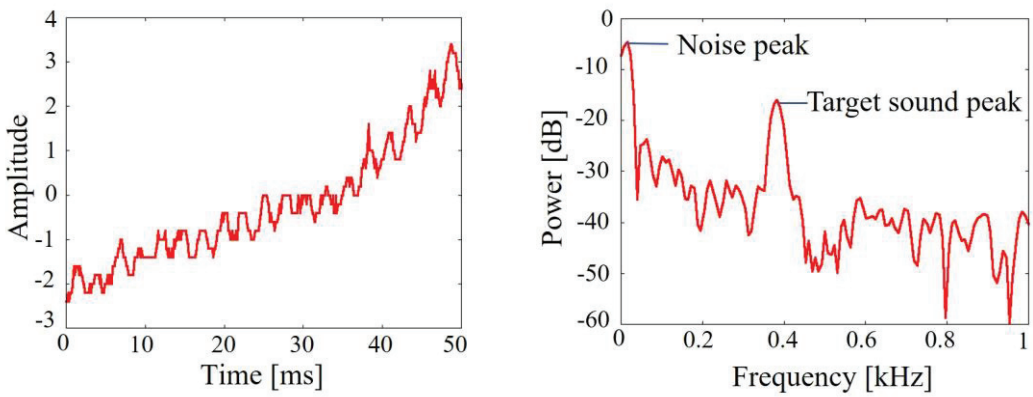


Figure 2 – Input image taken by using a CMOS image sensor.



(a) Waveform (b) Power spectrum

Figure 3 – Results of sound capture.

3. LOWER FREQUENCY NOISE REDUCTION

In this paper, we propose a noise reducing method for the captured sound by a digital filter considering the exposure-time of the CMOS image sensor. Figure 4 shows an image of a waveform of the captured sound. In the frequency analysis, the captured waveform is considered to be repeating as shown in Fig. 4. For this reason, the noise dependent on the frame is included. Therefore, this noise depends on the exposure-time of the sensor while taking an image. In addition, it shows that the lower frequency than the writing frequency based on exposure-time of the sensor cannot be captured by the sensor. Thus, we can reduce the noise by a digital filter considering exposure-time. Figure 5 shows the overview of the proposed method. In this paper, use a general high-pass filter of which cutoff frequency is determined by the exposure-time. The cutoff frequency of the high pass filter is given by Eq. (7).

$$F_{\mathrm{c}} = \frac{2}{T_{\mathrm{pict}}}, \tag{7}$$

$$T_{\mathrm{pict}} = H_{\mathrm{pict}} T_{\mathrm{lin}}, \tag{8}$$

where F_{c} is the target cutoff frequency, T_{pict} is the exposure-time of the sensor for an image, H_{pict} is vertical total numbers of pixels in an image, T_{lin} is the exposure-time of the sensor for each line. Using F_{c}, it is possible to design high-pass filter based on exposure-time and to reduce the lower frequency that cannot be captured with the sensor. The designed high-pass filter is convolved by Eq. (9) into $s(n)$ given in Eq. (5) to obtain an output signal $r(n)$.

$$r(n) = \sum_{m=0}^{N-1} h_{high}(m) * s(n-m), \tag{9}$$

where $h_{high}(n)$ is the impulse response of a high-pass filter designed with the cutoff frequency F_{c} given in Eq. (7) and N is the filter length.

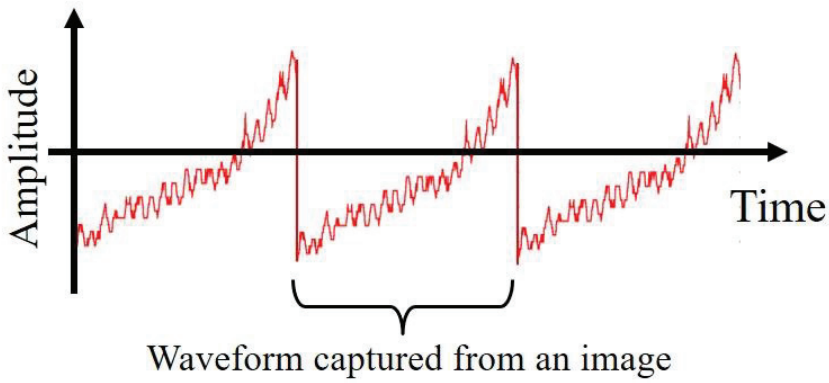


Figure 4 – An image of a waveform of the captured sound.

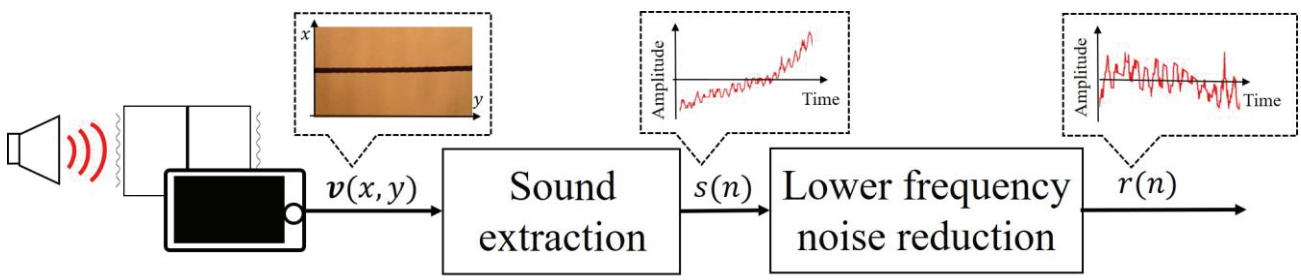


Figure 5 – Overview of the proposed method.

4. EVALUATION EXPERIMENT

4.1 Experimental Conditions

We carried out an experiment to verify that the lower frequency noise is reduced by the proposed method. Table 1 shows the measurement conditions. Figure 6 shows the equipment arrangement. We used an A4 paper printed a thin line in its center as the vibrating object. The vibrating object was located in 10 mm front of the loudspeaker under the light sources. In order to obtain enough illuminance, we used two incandescent bulbs. Pure tones with 0.1 kHz, 0.4 kHz and 0.8 kHz were respectively used as sound sources and emitted from the loudspeaker. The threshold value of Eq. (4) was set to 100 in this experiment. In this experiment, we compare spectral power to evaluate the performance for reduction of lower frequency noise.

4.2 Experimental Results

Figure 7 shows the frequency response of the high-pass filter used in the experiment. From Fig. 7, it can be seen that 0.02 kHz that is the frequency of noise at this experiment is reduced. The waveforms and power spectra of the captured sound by the original sound capture are shown in Fig. 8 and 9. As shown in Fig. 8 (a), the waveform of captured sound has an upward sloping shape. Also, from Fig. 8 (b), the lower frequency noise is mixed as shown in Fig. 9. The waveforms and power spectra of the captured sound by the proposed method are shown in Fig. 10 and 11. It can be seen from Fig. 10 and 11 that lower frequency noise is reduced by the proposed method compared to the original sound capture. By reducing the lower frequency noise, the spectral peak of the emitted pure tone becomes relatively prominent. Table 2 shows the signal-to-noise ratio (SNR) of the captured signals. Here, the signal is

Table 1 – Experimental conditions.

Environment	Soundproof room ($T_{60} = 100$ ms)
Background noise	$L_{\mathrm{A}} = 27.5$ dB
Sound sources	Pure tone (0.2, 0.4 and 0.8 kHz)
Sound pressure level	$L_a = 110$ dB at 0 m
Sampling frequency	8 kHz
Quantization bits	16 bits
Camera (Smartphone)	SONY, Xperia Z1
Size of image	3,840 × 2,160 pixels
Light source	Incandescent bulb (6,400 lm) × 2
Filter length	$N = 200$ samples
Exposure-time of the sensor	$T_{\mathrm{pict}} = 50$ ms

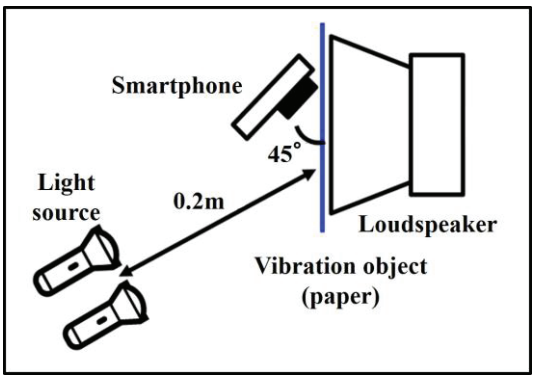


Figure 6 – Equipment arrangement (Top view).

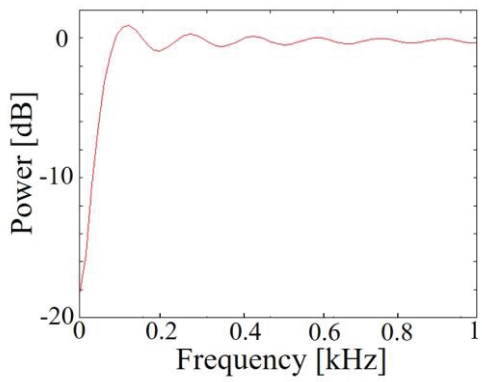
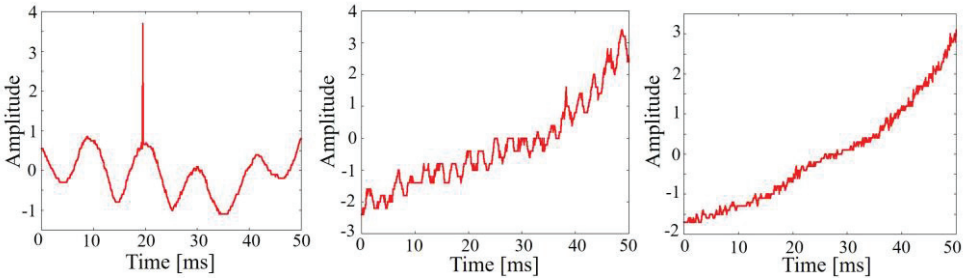


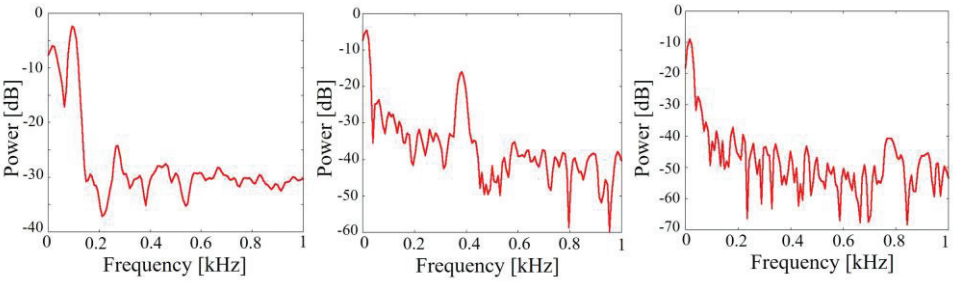
Figure 7 – Frequency response of the high-pass filter.

the emitted pure tone, and the noise is a signal related to the exposure-time of the CMOS image sensor. Here, we did not consider other frequency components to evaluate SNR. SNR is improved an average of 16.8 dB by using proposed method as shown Table 2. However, the noise still remains in the captured sound obtained by the proposed method. This is because the deflection of the paper is extracted as noise. Therefore, it is required to consider using an object other than paper as a vibrating object and reducing noise common to multiple frames.



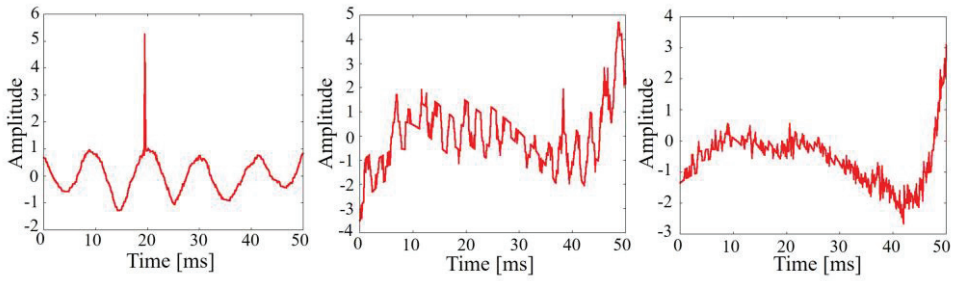
(a) 0.1 kHz (b) 0.4 kHz (c) 0.8 kHz

Figure 8 – Waveforms of original sound capture.



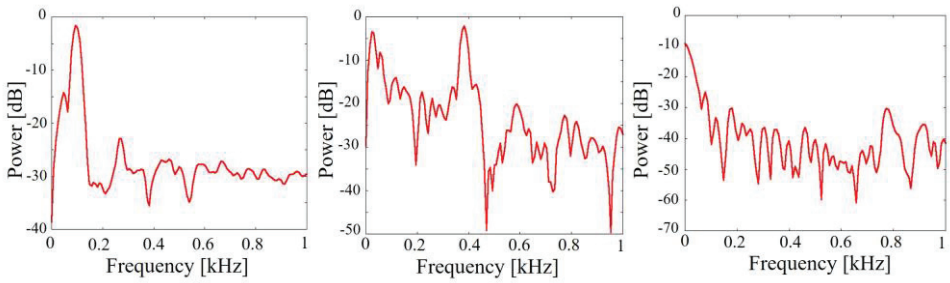
(a) 0.1 kHz (b) 0.4 kHz (c) 0.8 kHz

Figure 9 – Power spectra of original sound capture.



(a) 0.1 kHz (b) 0.4 kHz (c) 0.8 kHz

Figure 10 – Waveforms with proposed method.



(a) 0.1 kHz (b) 0.4 kHz (c) 0.8 kHz

Figure 11 – Power spectra with proposed method.

Table 2 – Evaluation result by SNR

Sound source	Original sound capture	Proposed method
0.1 kHz	3.6 dB	**22.1** dB
0.4 kHz	-11.4 dB	**4.3** dB
0.8 kHz	-34.8 dB	**-18.5** dB

5. CONCLUSIONS

We attempt to capture sound from an image with the CMOS image sensor. However, lower frequency noise is included in the captured sound using the CMOS image sensor due to the inclination and distortion of the object to be photographed. In this paper, to solve this problem, we proposed a noise reducing method for the captured sound by a digital filter considering the exposure-time of the CMOS image sensor. We carried out an experiment to verify that the lower frequency noise is reduced by the proposed method. The result of the evaluation experiment showed that SNR is improved an average of 16.8 dB by using proposed method. In the future, we will consider using an object other than paper as a vibrating object and reducing noise common to multiple frames.

ACKNOWLEDGEMENTS

This work was partly supported by JST-COI and JSPS KAKENHI Grant Numbers JP18K19829, JP19H04142, and R-GIRO (Ritsumeikan Global Innovation Research Organization) funded by Ritsumeikan University.

REFERENCES

1. M. A. Clark, "An acoustic lens as a directional microphone," *Journal of the Acoustical Society of America*, vol. 25, no. 4, p. 829, 1953.
2. A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," *Association for Computing Machinery Transactions on Graphics*, vol. 33, no. 4, pp. 79:1-79:10, 2014.
3. Y. Fuse, Y. Yasumi and T. Takiguchi, "Sound recovery using vibration modes of the object in a video," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 2027-2031, 2018.
4. M. Hua, L. Zhou, C. Liu and Z. Li, "The research of vibration detection using the visual microphone technology," *10th International Conference on Measuring Technology and Mechatronics Automation*, pp. 256-258, 2018.
5. S. Baker, E. Bennett, S. B. Kang and R. Szeliski, "Removing rolling shutter wobble," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2392-2399, 2010.
6. J. Chun, H. Jung and C. Kyung, "Suppressing rolling-shutter distortion of CMOS image sensors by motion vector detection," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 1479-1487, 2008.