APSIPA Transactions on Signal and Information Processing, 2025, 14, e5

Original Paper

Improvement of Sound Quality in Visual Microphone by Manipulation of Focused Area

Hayata Nakano*, Yuting Geng, Kenta Iwai and Takanobu Nishiura

Ritsumeikan University, Osaka, Japan

ABSTRACT

This paper presents methods for improving the sound quality of visual microphone by emphasizing the focused area in a captured video. When sound reaches an object, it causes vibrations of the object's surface. Thus, the visual microphone can extract sound by measuring the displacement of the object captured by the camera. In the captured video, there may be area that arise blurred (out-of-focused) due to the depth of field. In out-of-focused area of the captured video, it is difficult to accurately measure the displacement of the object being vibrated by sound, which may deteriorate the quality of the extracted sound. Here, the out-of-focused area is not taken into account in conventional sound extraction methods. In this paper, we propose three methods to extract sound by focusing on the focused area, where displacement can be measured more accurately than in the out-of-focused area. The proposed methods utilize out-of-focused area removal, weighted phase variation, and both processing to emphasize the measured displacement in the focused area of the captured video by using the focal rate that represents the degree of focus. Experimental results show that the proposed methods improve the quality of the extracted sound compared to the conventional method.

*Corresponding author: Hayata Nakano, is0516vr@ed.ritsumei.ac.jp

Received 22 November 2024; revised 23 January 2025; accepted 25 February 2025
ISSN 2048-7703; DOI 10.1561/116.20240087

Keywords: Sound extraction, captured video, sound quality improvement, out-of-focused area removal, weighted phase variation

1 Introduction

In sound recording, an air-conduction microphone is typically used to convert a sound wave into an electrical signal when a sound wave arrives at the diaphragm of the microphone. Consequently, the microphone acquires not only the target sound but also the background noise. The visual microphone [4, 14, 18, 19] was proposed to address this problem. Considering an object being vibrated by sound wave, it is able to acquire the sound by measuring the displacement of the object being vibrated by sound. The visual microphone acquires the sound from only visual information without the use of the audio data of the video. Based on the principle of the visual microphone, the sound signal extracted by the visual microphone contains little noise unless there are noise sources near the camera. Here, the extracted sound of the visual microphone is weaker at higher frequencies due to the effect of structural resonance characteristics. This is due to the fact that, the higher the frequency, the smaller the displacement and the greater the attenuation for most materials [4].

Many cameras for consumer products are equipped with rolling shutter image sensors [1, 3]. The rolling shutter image sensors in the camera sequentially exposes the pixel rows with a temporal offset between each row. Therefore, when capturing the object being vibrated by sound, rolling shutter distortion [1] arises in the captured images as a result of differences in exposure start times. Based on the rolling shutter distortion, the time variation of the object being vibrated by sound is measured for each row of the captured images [7, 2]. The visual microphone extracts the sound signal by measuring the displacement of the object in each row of the captured images using this distortion. Here, we consider the case where out-of-focused (i.e., blurred) area arises in the captured video. Therefore, when extracting sound including out-of-focused area in the captured video, it is difficult to accurately measure the displacement of the object being vibrated by sound, leading to degradation in the quality of the extracted sound signal. However, out-of-focused area is not taken into account in conventional sound extraction methods [4, 14, 18, 19].

We proposed sound extraction methods to improve the sound quality of visual microphone, based on our preliminary studies [11, 10]. We focus on the out-of-focused area in captured video and we define a measure "focal rate" to represent the degree of focus. Based on the focal rate, we remove out-of-focused area or apply weighted emphasis to the focused area to enhance the quality of the extracted sound [11]. Additionally, by combining these two methods, we aim to improve the quality of the extracted sound by further

emphasising the focused area [10]. In [11, 10], the effectiveness of the proposed methods was confirmed through simple experiments. To validate the effectiveness of the proposed methods, we conducted sound extraction experiments under several conditions.

This paper is organized as follows. In Section 2, we explain the principle and problem with the conventional sound extraction method [4] for visual microphone. Then, the proposed sound extraction methods are explained in Section 3, and experimental results are shown in Section 4. Finally, conclusions and future work are shown in Section 5.

2 Conventional sound extraction method in visual microphone

This Section explains the method of sound extraction using video captured by a rolling shutter camera [4]. Previous research [4] primarily focused on sound extraction by using high-speed cameras, and the method based on phase variation in the complex steerable pyramid [13, 12, 17] was proposed. Also, the method using a rolling shutter camera based on the above approach was proposed specifically for videos with lower frame rates.

An overview of this method is shown in Figure 1. First, a rolling shutter camera captures the displacements of an object being vibrated by sound and a captured RGB video is converted to a grayscale video $I(x, y, n)$, where x and y are the column and row indices, and n is the frame index. Then, the sound signal is obtained based on the phase variation of the captured video $I(x, y, n)$. The phase variation is calculated using a complex steerable pyramid [13, 12, 17] for the captured video $I(x, y, n)$. The complex steerable pyramid is composed of sub-band images for each scale r. $S_r(x, y, n)$, which is the decomposition of $I(x, y, n)$, is given by

$$S_r(x, y, n) = A_r(x, y, n)\, \mathrm{e}^{\mathrm{j}\phi_r(x,y,n)}, \tag{1}$$

where j, $A_r(x, y, n)$ and $\phi_r(x, y, n)$ denote the imaginary unit, the amplitude of the sub-band video and its phase, respectively. Then, the number of columns W_r and rows H_r in the captured video $S_r(x, y, n)$ for each scale r are given by

$$W_r = 2^{-r} W, \tag{2}$$

$$H_r = 2^{-r} H, \tag{3}$$

where W and H denote the number of columns and rows in $I(x, y, n)$. By using the phase $\phi_r(x, y, n)$ of each row in each video frame, the sound signal can be extracted, as the vibration of the object causes the phase variation [16,

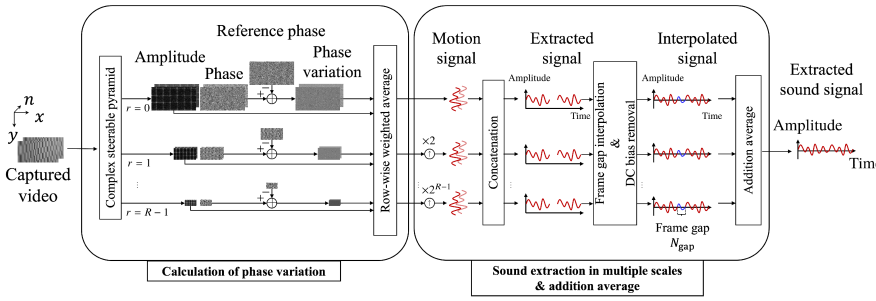


Figure 1: Overview of conventional sound extraction method for visual microphone.

6, 5]. The phase variation $\phi_{\mathrm{v},r}(x, y, n)$ is calculated from the reference frame $\phi_r(x, y, n_0)$ as

$$\phi_{\mathrm{v},r}(x, y, n) = \phi_r(x, y, n) - \phi_r(x, y, n_0), \tag{4}$$

where n_0 denotes the index representing the reference frame. Then, a row-wise weighted average $\Phi_r(y, n)$ of the phase variation $\phi_{\mathrm{v},r}(x, y, n)$ is calculated by

$$\Phi_r(y, n) = \frac{1}{W_r} \sum_{x=0}^{W_r-1} A^2(r, x, y, n)\phi_{\mathrm{v},r}(x, y, n), \tag{5}$$

Here, the amplitude $A_r(x, y, n)$ is used as the weight since the phase can be accurately estimated for pixels with large amplitude. This results in one value per row, representing the amplitude of the sound signal. As the length of $\Phi_r(y, n)$ is H_r for each r, the signal length H_r is aligned to H through upsampling by a factor of 2^r for each r. Here, $\Phi_r(y, n)$ for each r is a two-dimensional signal. Therefore, the averaged phase variation is transformed to a one-dimensional signal $\tilde{\Phi}_r(t)$ for each r, as

$$\tilde{\Phi}_r(t) = \Phi_r(y, n), \tag{6}$$

$$t = y + (H + N_{\mathrm{gap}})n, \tag{7}$$

where t and N_{gap} denote the time index and the number of samples in the frame gap. The frame gaps arise due to the characteristics of rolling shutter image sensors [1]. Therefore, the autoregressive model [8] is utilized to interpolate the frame gaps in the extracted signal $\tilde{\Phi}_r(t)$ to obtain the interpolated signal $d_r(t)$.

The extracted signal $\tilde{\Phi}_r(t)$ may contain direct current (DC) bias, which is removed as

$$d_r(t) = \tilde{\Phi}_r(t) - \bar{\Phi}_r, \tag{8}$$

where $\bar{\Phi}_r$ is the average amplitude in the extracted signal for each scale r and $d_r(t)$ is the signal after removal of the DC bias. Finally, a sound signal $u(t)$ is obtained by calculating sum average of $d_r(t)$ with respect to r by

$$u(t) = \frac{1}{R} \sum_{r=0}^{R-1} d_r(t), \tag{9}$$

where R is the number of scales of the complex steerable pyramid.

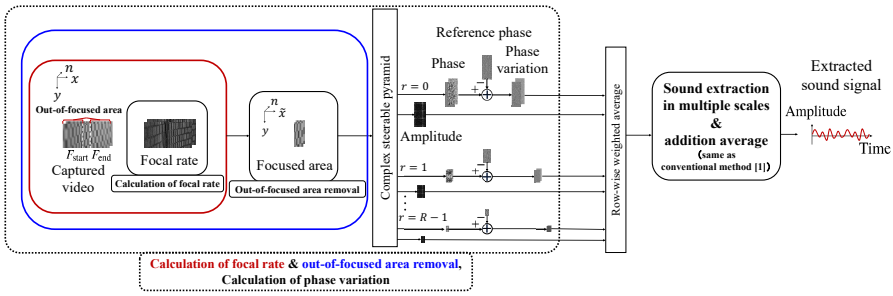
The conventional method performs sound extraction without handling out-of-focused area in the captured video. However, the video captured by the camera may contain out-of-focused area due to the effect of depth of field, and the edges of the video images are blurred. Therefore, the conventional method cannot accurately measure the displacement in the out-of-focused area because it uses the phase related to the edge, which may cause the sound quality of the extracted sound to deteriorate.

3 Proposed sound extraction method for visual microphone

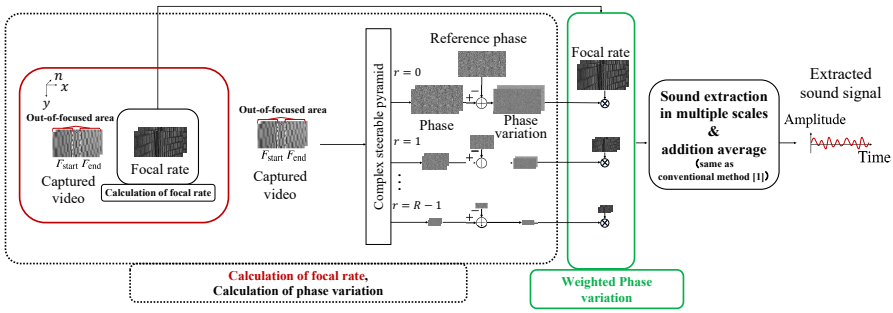
We propose methods for improving the sound quality of extracted sound by utilizing the focused area to solve the problem described in the previous section. In this paper, we consider that the out-of-focused area of the captured video occurs in the column direction (vertical direction) to simplify the discussion. Our proposed sound extraction methods aim to improve the sound quality of the extracted sound by emphasizing the focused area of the captured video. The proposed methods emphasize the displacement measured from the focused area of the captured video based on the focal rate. Here, we introduce three proposed methods utilize out-of-focused area removal, weighted phase variation, and both processing to emphasize the measured displacement in the focused area of the captured video. The procedure of three proposed methods is illustrated in Figure 2. Section 3.1 describes the processing procedure in Figure 2 (a). Section 3.2 describes the processing procedure in Figure 2 (b). Finally, Section 3.3 describes the processing procedure in Figure 2 (c).

3.1 Out-of-focused area removal

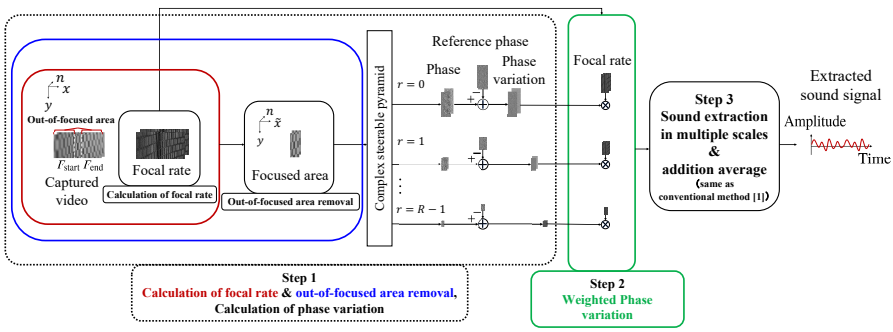
We describe a method in Figure 2 (a) for improving the sound quality of the extracted sound by removing the out-of-focused area in the captured video using the focal rate. A schematic diagram of the removal of out-of-focused area in the captured video is shown in Figure 3. As mentioned above, when out-of-focused area of the captured video is used for sound extraction, the phase variation cannot be calculated correctly and the quality of the extracted sound signal degrades. Therefore, we only utilize the focused area to extract



(a) Out-of-focused area removal



(b) Weighted phase variation



(c) Combination of out-of-focused area removal and weighted phase variation

Figure 2: Overview of proposed sound extraction methods for visual microphone.

the sound. Here, in the out-of-focused area, the edges of the captured video are blurred, whereas the edges are relatively clear in the focused area. Therefore, the focal rate is calculated using a Sobel filter, which is commonly employed as an edge enhancement filter [15]. Here, the sound extraction methods in this paper are only use the displacement in the horizontal direction; thus, a

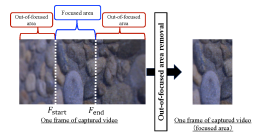


Figure 3: Overview of out-of-focused area removal.

horizontal Sobel filter is used. In out-of-focused area, the change in pixel values is smooth, resulting in small absolute values of horizontal gradients. Meanwhile, in focused area, the change in pixel values is sharp, resulting in larger absolute values of horizontal gradients. Therefore, we define the absolute values of horizontal gradients as the focal rate.

The focal rate $M(x, y, n)$ for each frame of the captured video $I(x, y, n)$ is calculated as

$$M(x, y, n) = |K(x, y) * I(x, y, n)|, \tag{10}$$

where $*$ denotes two dimensional convolution operater and $K(x, y)$ denotes a 3Œ3 horizontal Sobel filter. We assume that the focused area does not change significantly between frames of the captured video. Therefore, we remove the out-of-focused area by using the focal rate $M(x, y, 0)$ calculated from the first frame. Then, the focal rate $M(x, y, 0)$ is denoted as $M(x, y)$.

The column-wise mean of $\overline{M}(x)$ is calculated as

$$\overline{M}(x) = \frac{1}{H} \sum_{y=0}^{H-1} |M(x, y)|, \tag{11}$$

Then, we define $\boldsymbol{F}$ as the set of column indices x such that $\overline{M}(x)$ is larger than the threshold value $\overline{M}_{\text{th}}$:

$$\boldsymbol{F} = \{x \in \mathbb{N} \mid x \in [0, W-1],\ \overline{M}(x) \geq \overline{M}_{\text{th}}\}, \tag{12}$$

$$\overline{M}_{\text{th}} = \frac{\alpha}{W} \sum_{x=0}^{W-1} \overline{M}(x),\ \alpha > 0, \tag{13}$$

where $\overline{M}_{\text{th}}$ is calculated as α times the row-wise means $\overline{M}(x)$. The effect of removing out-of-focused area can be amplified by increasing this α. The column indices indicating the start point F_{start} and endpoint F_{end} of the focused area are calculated as

$$(F_{\text{start}}, F_{\text{end}}) = \left(\min_{x \in \boldsymbol{F}} x,\ \max_{x \in \boldsymbol{F}} x \right). \tag{14}$$

We remove the out-of-focused area from the captured video $I(x, y, n)$ and obtain $I_{\text{focal}}(\tilde{x}, y, n)$ which is the video of the focused area as

$$I_{\text{focal}}(\tilde{x}, y, n) = I(\tilde{x} + F_{\text{start}} - 1, y, n), 0 \leq \tilde{x} \leq \tilde{W} - 1, \tag{15}$$

$$\tilde{W} = F_{\text{end}} - F_{\text{start}} + 1, \tag{16}$$

where $\tilde{W}$ denotes the number of columns in $I_{\text{focal}}(\tilde{x}, y, n)$.

Here, the sound is extracted from the captured video $I_{\text{focal}}(\tilde{x}, y, n)$. As in Section 2, phase variation is calculated based on the complex steerable pyramid. The number of columns of captured video is $\tilde{W}$ in this method, whereas it was W in Section 2. Therefore, $S_r(\tilde{x}, y, n)$, which is the decomposition of $I(\tilde{x}, y, n)$, is given by

$$S_r(\tilde{x}, y, n) = A_r(\tilde{x}, y, n)\, \mathrm{e}^{\mathrm{j}\phi_r(\tilde{x}, y, n)}, \tag{17}$$

Then, the numbers of columns $\tilde{W}_r$ and rows H_r in the captured video $S_r(\tilde{x}, y, n)$ for each scale r are given by

$$\tilde{W}_r = 2^{-r}\tilde{W}, \tag{18}$$

$$H_r = 2^{-r}H. \tag{19}$$

The phase variation $\phi_{\text{v},r}(\tilde{x}, y, n_0)$ from the reference frame $\phi_r(\tilde{x}, y, n_0)$ is calculated as

$$\phi_{\text{v},r}(\tilde{x}, y, n) = \phi_r(\tilde{x}, y, n) - \phi_r(\tilde{x}, y, n_0). \tag{20}$$

We calculate the row-wise weighted average $\Phi_{\text{rem},r}(y, n)$ of the phase variation $\phi_{\text{v},r}(\tilde{x}, y, n)$. The row-wise weighted average $\Phi_{\text{rem},r}(y, n)$ is calculated as

$$\Phi_{\text{rem},r}(y, n) = \frac{1}{\tilde{W}_r} \sum_{\tilde{x}=0}^{\tilde{W}_r - 1} A_r^2(\tilde{x}, y, n)\phi_{\text{v},r}(\tilde{x}, y, n), \tag{21}$$

As the length of $\Phi_{\text{rem},r}(y, n)$ is H_r for each r, the signal length H_r is aligned to H through upsampling by a factor of 2^r for each r.

Henceforth, $\Phi_r(y, n)$ in the conventional method described in Section 2 is changed to $\Phi_{\text{rem},r}(y, n)$, and the sound signal is extracted by the process similar to the conventional method. $\Phi_{\text{rem},r}(y, n)$ for each r is a two-dimensional signal. Therefore, the averaged phase variation is transformed to a one-dimensional signal $\tilde{\Phi}_{\text{rem},r}(t)$ for each r, as

$$\tilde{\Phi}_{\text{rem},r}(t) = \Phi_{\text{rem},r}(y, n). \tag{22}$$

As in Section 2, an autoregressive model [8] is utilized to interpolate the frame gaps in the extracted signal $\tilde{\Phi}_{\text{rem},r}(t)$ to obtain an interpolated signal $d_{\text{rem},r}(t)$. The DC bias is removed from the extracted signal $\tilde{\Phi}_{\text{rem},r}(t)$ for each r as

$$d_{\text{rem},r}(t) = \tilde{\Phi}_{\text{rem},r}(t) - \bar{\Phi}_{\text{rem},r}, \tag{23}$$

where $\bar{\Phi}_{\mathrm{rem},r}$ is the average amplitude in the extracted signal for each scale r and $d_{\mathrm{rem},r}(t)$ is the signal after removal of the DC bias. Finally, a sound signal $u_{\mathrm{rem}}(t)$ is obtained by calculating sum average of $d_{\mathrm{rem},r}(t)$ with respect to r as

$$u_{\mathrm{rem}}(t) = \frac{1}{R}\sum_{r=0}^{R-1} d_{\mathrm{rem},r}(t). \tag{24}$$

3.2 Weighted phase variation

We describe a method in Figure 2 (b) to improve the sound quality of the extracted sound by weighting the phase variation according to the focal rate of the captured video.

First, as in the previous section, the focal rate $M(x, y, n)$ is calculated for each frame of the captured video $I(x, y, n)$ by (10). As in Section 2, the phase variation $\phi_{\mathrm{v},r}(x, y, n)$ is calculated based on the complex steerable pyramid. $\phi_{\mathrm{v},r}(x, y, n)$ is given by (1)–(4). The phase variation $\phi_{\mathrm{v},r}(x, y, n)$ is weighted with the focal rate $M_r(x, y, n)$, which can be given as :

$$\Phi_{\mathrm{weight},r}(y, n) = \frac{1}{W_r}\sum_{x=0}^{W_r-1} M_r^2(x, y, n)\phi_{\mathrm{v},r}(x, y, n), \tag{25}$$

The focal rate $M_r(x, y, n)$ for each scale r is calculated by downsampling the $M(x, y, n)$ to 2^{-r} times the resolution. As the length of $\Phi_{\mathrm{weight},r}(y, n)$ is H_r for each r, the signal length H_r is aligned to H through upsampling by a factor of 2^r for each r.

Henceforth, $\Phi_r(y, n)$ in the conventional method described in Section 2 is changed to $\Phi_{\mathrm{weight},r}(y, n)$, and the sound signal is extracted by the process similar to the conventional method. $\Phi_{\mathrm{weight},r}(y, n)$ for each r is a two-dimensional signal. Therefore, the averaged phase variation is transformed to a one-dimensional signal $\tilde{\Phi}_{\mathrm{weight},r}(t)$ for each r, as

$$\tilde{\Phi}_{\mathrm{weight},r}(t) = \Phi_{\mathrm{weight},r}(y, n). \tag{26}$$

As in Section 2, an autoregressive model [8] is utilized to interpolate the frame gaps in the extracted signal $\tilde{\Phi}_{\mathrm{weight},r}(t)$ to obtain an interpolated signal $d_{\mathrm{weight},r}(t)$. The DC bias is removed from the extracted signal $\tilde{\Phi}_{\mathrm{weight},r}(t)$ for each r as

$$d_{\mathrm{weight},r}(t) = \tilde{\Phi}_{\mathrm{weight},r}(t) - \bar{\Phi}_{\mathrm{weight},r}, \tag{27}$$

where $\bar{\Phi}_{\mathrm{weight},r}$ is the average amplitude in the extracted signal for each scale r and $d_{\mathrm{weight},r}(t)$ is the signal after removal of the DC bias.

Finally, a sound signal $u_{\mathrm{weight}}(t)$ is obtained by calculating sum average of $d_{\mathrm{weight},r}(t)$ with respect to r as

$$u_{\mathrm{weight}}(t) = \frac{1}{R}\sum_{r=0}^{R-1} d_{\mathrm{weight},r}(t). \tag{28}$$

3.3 Combination of out-of-focused area removal and weighted phase variation.

The method in Figure 2 (c) is based on a combination of two methods, out-of-focused area removal described in Section 3.1 and weighted phase variation described in Section 3.2. By combining these two methods, it may be possible to enhance the displacement measured in the focused area of the captured video.

Step 1: Calculation of focal rate and out-of-focused area removal

This step calculates the focal rate and removes out-of-focused area in the captured video as described in Section 3.1. The focal rate $M(x, y, n)$ is calculated for each frame of the captured video by (10). Additionally, the out-of-focused area of the captured video is removed based on the focal rate $M(x, y, n)$ by (11)–(16).

The phase variation $\phi_{\mathrm{v},r}(\tilde{x}, y, n)$ for each scale is calculated from the captured video of the focused area by (17)–(20).

Step 2: Weighted phase variation

Here, the focal rate $M(x, y, n)$ and phase variation $\phi_{\mathrm{v},r}(\tilde{x}, y, n)$ have different sizes. Therefore, the focal rate $M_{\mathrm{focal}}(\tilde{x}, y, n)$ of the focal area is calculated as

$$M_{\mathrm{focal}}(\tilde{x}, y, n) = M(\tilde{x} + F_{\mathrm{start}}, y, n),\ 0 \le \tilde{x} \le \tilde{W} - 1, \tag{29}$$

The phase variation $\phi_{\mathrm{v}}(r, x, y, n)$ is weighted with the focal rate $M_{\mathrm{focal}}(\tilde{x}, y, n)$ as

$$\Phi_{\mathrm{prop},r}(y, n) = \frac{1}{\tilde{W}_r}\sum_{\tilde{x}=0}^{\tilde{W}_r-1} M^2_{\mathrm{focal},r}(\tilde{x}, y, n)\phi_{\mathrm{v},r}(\tilde{x}, y, n), \tag{30}$$

The focal rate $M_{\mathrm{focal},r}(\tilde{x}, y, n)$ for each scale r is calculated by downsampling the $M_{\mathrm{focal}}(\tilde{x}, y, n)$ to 2^{-r} times the resolution.

Step 3: Extraction of sound in multiple scales

$\Phi_r(y, n)$ in the conventional method described in Section 2 is changed to $\Phi_{\mathrm{prop},r}(y, n)$, and the sound signal is extracted by the process similar to the conventional method. $\Phi_{\mathrm{prop},r}(y, n)$ for each r is a two-dimensional signal. Therefore, the averaged phase variation is transformed to a one-dimensional signal $\tilde{\Phi}_{\mathrm{prop},r}(t)$ for each r, as

$$\tilde{\Phi}_{\mathrm{prop},r}(t) = \Phi_{\mathrm{prop},r}(y, n). \tag{31}$$

As in Section 2, an autoregressive model [8] is utilized to interpolate the frame gaps in the extracted signal $\tilde{\Phi}_{\mathrm{prop},r}(t)$ to obtain an interpolated signal $d_{\mathrm{prop},r}(t)$. The DC bias is removed from the extracted signal $\tilde{\Phi}_{\mathrm{prop},r}(t)$ for each r as

$$d_{\mathrm{prop},r}(t) = \tilde{\Phi}_{\mathrm{prop},r}(t) - \bar{\Phi}_{\mathrm{prop},r}, \tag{32}$$

where $\bar{\Phi}_{\mathrm{prop},r}$ is the average amplitude in the extracted signal for each scale r and $d_{\mathrm{prop},r}(t)$ is the signal after removal of the DC bias.

Finally, a sound signal $u_{\mathrm{prop}}(t)$ is obtained by calculating sum average of $d_{\mathrm{prop},r}(t)$ with respect to r as

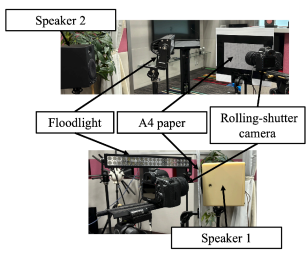
$$u_{\mathrm{prop}}(t) = \frac{1}{R}\sum_{r=0}^{R-1} d_{\mathrm{prop},r}(t). \tag{33}$$

4 Evaluation experiments

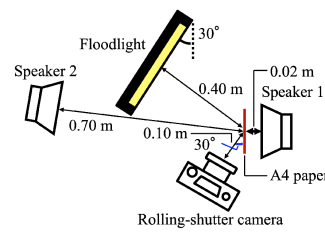
This Section describes the evaluation experiments conducted to confirm the effectiveness of the proposed method described in Section 3. Section 4.1 describes the experimental conditions of the sound extraction experiments, and Section 4.2 describes the experimental results and discussion of the sound extraction experiments using sine waves. Finally, Section 4.3 describes comparison of execution time between conventional and proposed methods.

4.1 Experimental conditions for sound extraction

Figures 4 and 5 show the experimental setup and equipment arrangement and printed patterns on A4 paper used as the object being vibrated by sound. Moreover, Tables 1 and 2 show the experimental conditions and equipment.

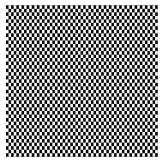


(a) Experimental scene

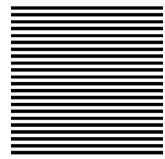


(b) Experimental equipment

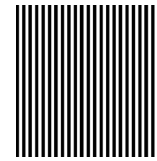
Figure 4: Experimental conditions for sound extraction.



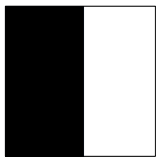
(a) Printed pattern A



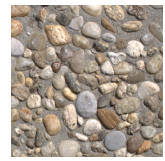
(b) Printed pattern B



(c) Printed pattern C



(d) Printed pattern D



(e) Printed pattern E

(f) Printed pattern F

Figure 5: Patterns to print on A4 paper.

Table 1: Experimental condition.

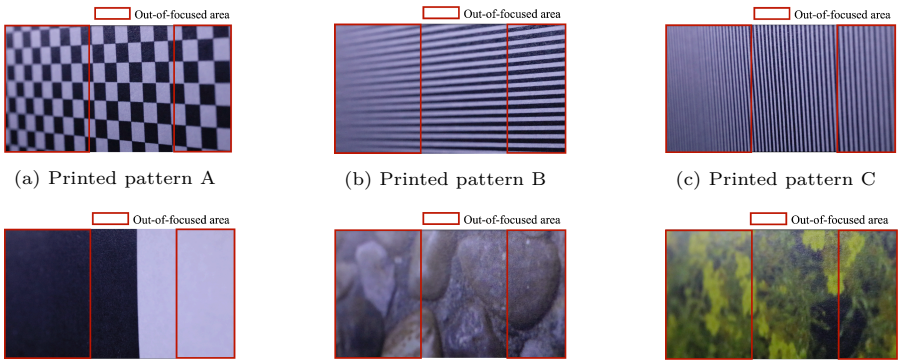
Ambient noise level	L_{A}=38.6 dB
Temperature / Humidity	21.3°C/20.9%
Sound source	Sine wave (300, 500, ⋯, 1,500 Hz)
Sound pressure level	85 dB at 0 m from A4 paper
Noise sound source	White noise
Noise sound pressure level	60 dB at 0 m from A4 paper
Sampling frequency / Quantization	8,000 Hz / 16 bits
Resolution of captured video (width × height)	1,920 × 1,080 px
Frame rate of captured video	60 fps
Exposure time of camera	1/4,000 s

Table 2: Experimental equipment.

Camera	Canon EOS 5D MarkIV
Camera lens	Canon MP-E 65 mm f/2.8 1-5x
Speaker 1	FOSTEX FE83En
Loudspeaker amplifier	BOSE 1705II
Speaker 2	YAMAHA MSP3
Floodlight	GOODGOODS GDGDS-WL02 (10,000 lm)

As shown in Figure 4, the A4 paper placed in front of the loudspeaker was captured with the rolling shutter camera, and the sound signal was extracted from the captured video. In order to evaluate the differences in the printed pattern of the A4 paper, the patterns shown in Figure 5 (a)–(f) were printed on the A4 paper, and experiments were conducted on each of them. In particular, Figure 5 (e), (f) is referenced from the Salzburg Texture Image Database (STex) [9]. Printed pattens A–D were chosen to evaluate the impact of vertical edges of printed pattern on the sound extraction accuracy of each method. Sound extraction from a captured video of rolling shutter camera measures

horizontal displacements. In this process, vertical edges are considered important for the accuracy of sound extraction. Furthermore, print patterns E and F feature stone and flower designs commonly seen daily. To obtain the video with out-of-focused area, the A4 paper was captured at 30°. This takes into account that capturing the object at this angle causes out-of-focused area in the video resulting from the relationship between focal length and depth of field. In addition, we used a floodlight for illumination so that we could accurately capture the A4 paper's vibration when the shutter speed was high. Figure 6 shows one frame of the captured video.



(a) Printed pattern A (b) Printed pattern B (c) Printed pattern C

(d) Printed pattern D (e) Printed pattern E (f) Printed pattern F

Figure 6: One frame of captured video for each printed pattern.

We used sine waves as the sound source. Additionally, each sine wave was radiated from the loudspeaker (speaker 1 in Figure 4) for the duration of 5 seconds. Here, as described in the introduction, the sound extracted by the visual microphone contains little noise unless there are noise sources near the camera. To prove this, we conducted an experiment using white noise. For printed pattern A, we compared the sound quality of the extracted sound from the captured video. White noise was radiated from the loudspeaker (speaker 2 in Figure 4). In this experiment, we set $n_0 = 0$ as the reference frame and $r_0 = 0$ as the reference scale. The scale of the complex steerable pyramid is set to $R = 2$. Table 3 shows the number of rows of video for each α in (13). Here, when applying the Complex steerable pyramid to a captured video, scale R determines the minimum resolution of the captured video from which sound can be extracted. Therefore, if α is increased and the number of columns is reduced too much, the complex steerable pyramid cannot be applied and sound cannot be extracted. The range of α in which sound could be extracted was 0.5 to 1.5 for printed patterns C and D, 0.5 to 2.0 for printed patterns A and B, 0.5 to 2.5 for printed pattern F, and 0.5 to 3.0 for printed pattern E.

Table 3: Number of columns of video in α.

Printed pattern	α in Eq. (13)					
	0.5	1.0	1.5	2.0	2.5	3.0
A	1,918	1,916	1,538	504	93	0
B	1,914	1,136	244	107	0	0
C	1,508	1,061	646	60	0	0
D	1,722	1,385	238	42	40	39
E	1,914	516	312	235	190	150
F	1,917	688	274	195	110	0

Henceforth, all methods of sound extraction are referred to as follows:

Conv : the method [4] described in Section 2.
Prop_{rem} : the method of sound extraction by removing out-of-focused area described in Section 3.1.
Prop_{wei} : the method of sound extraction by weighting the focal area described in Section 3.2.
Prop_{com} : the method described in Section 3.3.

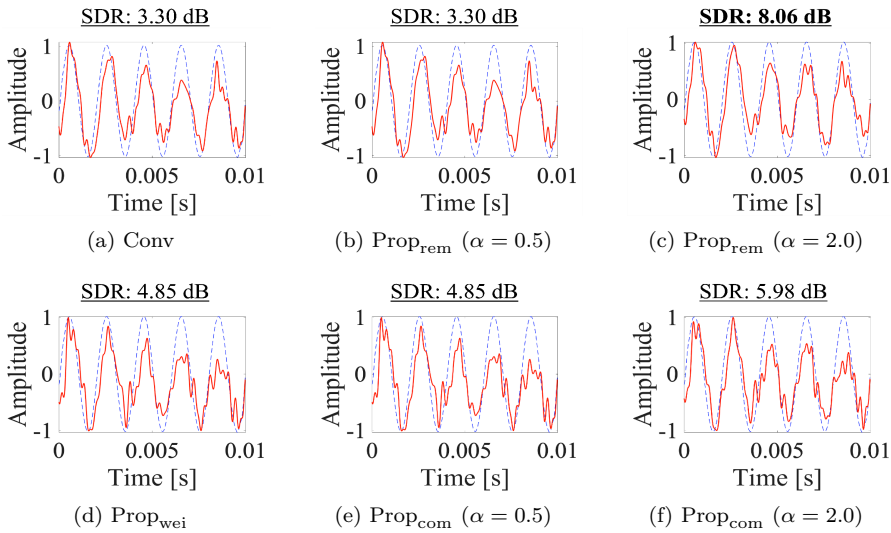
4.2 Experimental results of sound extraction

Figures 7–12 show the time waveforms of the 500 Hz sine waves extracted by each extraction method in each printed pattern. The SDR values for each time waveform are also shown in Figures 7–12. The SDR is calculated as

$$\text{SDR} = 10\log_{10}\left[\frac{\sum_{t=0}^{T-1} u_{\text{c}}^2(t)}{\sum_{t=0}^{T-1}\{u_{\text{c}}(t) - \lambda u(t)\}^2}\right], \tag{34}$$

$$\lambda = \sqrt{\frac{\sum_{t=0}^{T-1} u_{\text{c}}^2(t)}{\sum_{t=0}^{T-1} u^2(t)}}, \tag{35}$$

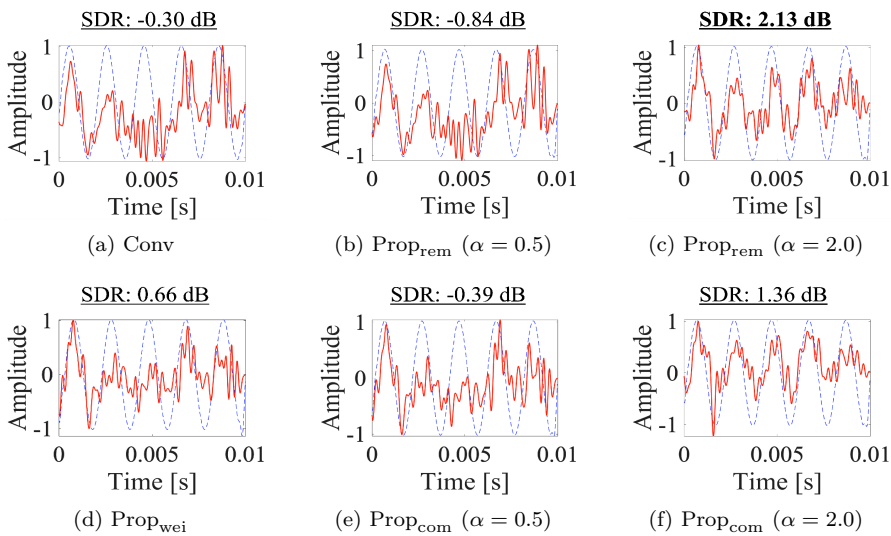
where $u_{\text{c}}(t)$ denotes the sine wave to be used as the sound source. Additionally, we experimented with $T = 80$ samples (0.01 s under the sampling frequency of 8,000 Hz). The time delay was adjusted by using cross-correlation so that $u(t)$ and $u_{\text{c}}(t)$ were in phase.



(a) Conv (b) Prop_{rem} ($\alpha = 0.5$) (c) Prop_{rem} ($\alpha = 2.0$)

(d) Prop_{wei} (e) Prop_{com} ($\alpha = 0.5$) (f) Prop_{com} ($\alpha = 2.0$)

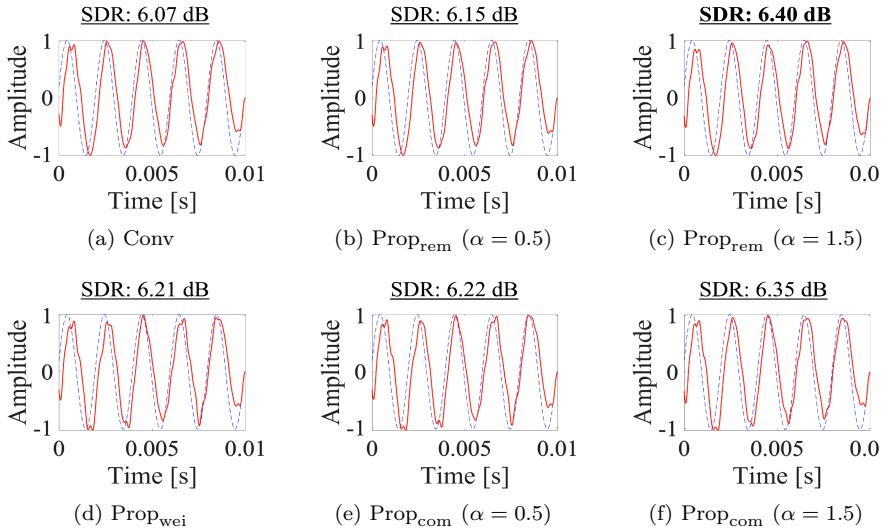
Figure 7: Time waveforms (500 Hz) of sound extracted by each method with printed pattern A (red line: extracted sound, blue dotted line: sound source).



(a) Conv (b) Prop_{rem} ($\alpha = 0.5$) (c) Prop_{rem} ($\alpha = 2.0$)

(d) Prop_{wei} (e) Prop_{com} ($\alpha = 0.5$) (f) Prop_{com} ($\alpha = 2.0$)

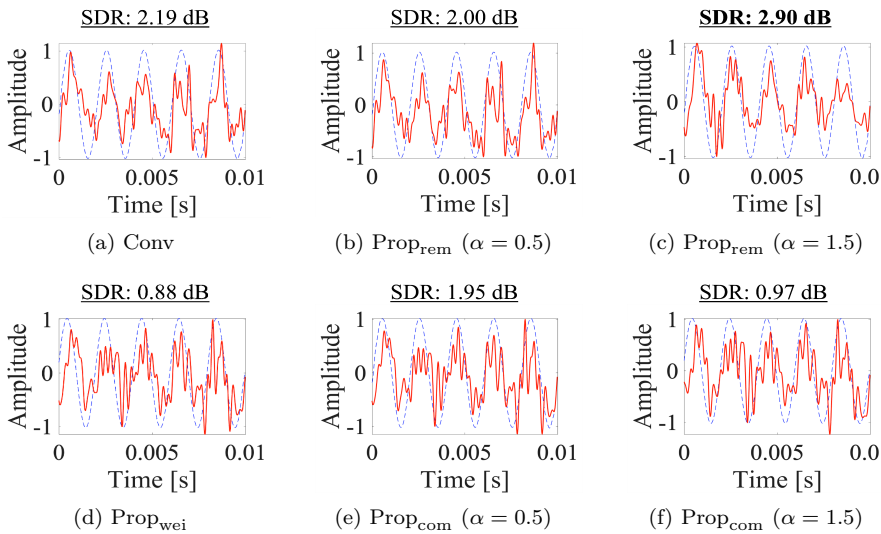
Figure 8: Time waveforms (500 Hz) of sound extracted by each method with printed pattern B (red line: extracted sound, blue dotted line: sound source).



(a) Conv (b) Prop_{rem} ($\alpha = 0.5$) (c) Prop_{rem} ($\alpha = 1.5$)

(d) Prop_{wei} (e) Prop_{com} ($\alpha = 0.5$) (f) Prop_{com} ($\alpha = 1.5$)

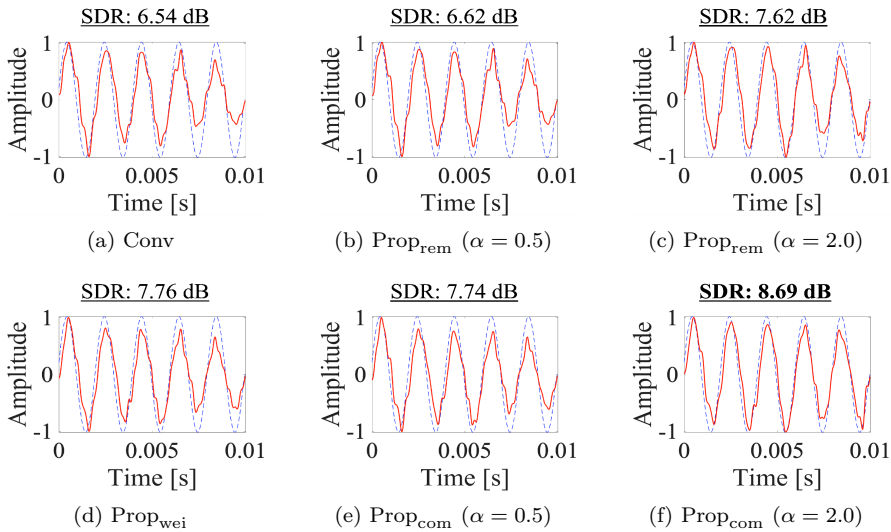
Figure 9: Time waveforms (500 Hz) of sound extracted by each method with printed pattern C (red line: extracted sound, blue dotted line: sound source).



(a) Conv (b) Prop_{rem} ($\alpha = 0.5$) (c) Prop_{rem} ($\alpha = 1.5$)

(d) Prop_{wei} (e) Prop_{com} ($\alpha = 0.5$) (f) Prop_{com} ($\alpha = 1.5$)

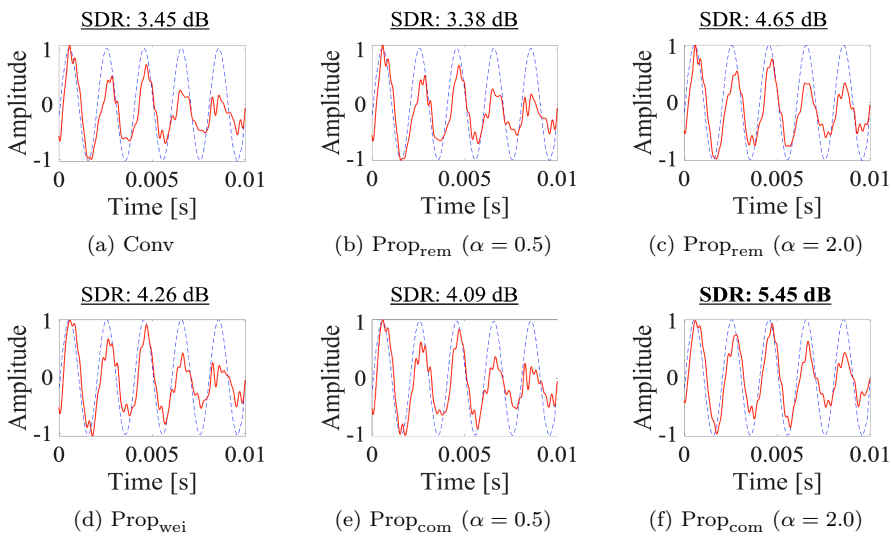
Figure 10: Time waveforms (500 Hz) of sound extracted by each method with printed pattern D (red line: extracted sound, blue dotted line: sound source).



(a) Conv (b) Prop_{rem} ($\alpha = 0.5$) (c) Prop_{rem} ($\alpha = 2.0$)

(d) Prop_{wei} (e) Prop_{com} ($\alpha = 0.5$) (f) Prop_{com} ($\alpha = 2.0$)

Figure 11: Time waveforms (500 Hz) of sound extracted by each method with printed pattern E (red line: extracted sound, blue dotted line: sound source).



(a) Conv (b) Prop_{rem} ($\alpha = 0.5$) (c) Prop_{rem} ($\alpha = 2.0$)

(d) Prop_{wei} (e) Prop_{com} ($\alpha = 0.5$) (f) Prop_{com} ($\alpha = 2.0$)

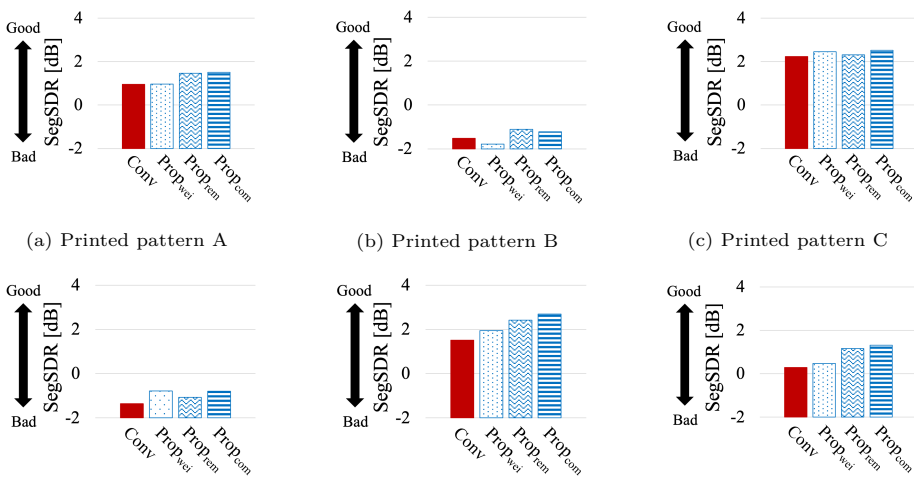
Figure 12: Time waveforms (500 Hz) of sound extracted by each method with printed pattern F (red line: extracted sound, blue dotted line: sound source).

Tables 5–10 show the results of the SegSDR (segmental signal-to-distortion ratio) evaluation of the sound quality of the extracted sound for each printed pattern shown in Figure 5. The values in Tables 5–10 indicate the SegSDR. The SegSDR is calculated as

$$\mathrm{SegSDR} = \frac{1}{K}\sum_{k=0}^{K-1} 10\log_{10}\left[\frac{\sum_{t=2kT}^{2kT+T-1} u_{\mathrm{c}}^2(t)}{\sum_{t=2kT}^{2kT+T-1}\{u_{\mathrm{c}}(t)-\lambda_k u(t)\}^2}\right], \tag{36}$$

$$\lambda_k = \sqrt{\frac{\sum_{t=2kT}^{2kT+T-1} u_{\mathrm{c}}^2(t)}{\sum_{t=2kT}^{2kT+T-1} u^2(t)}}, \tag{37}$$

where k denotes the segment index, K denotes the number of segments. As with the calculation of SegSDR, we experimented with $T = 80$ and $K = 50$. The time delay was adjusted by using cross-correlation so that $u(t)$ and $u_{\mathrm{c}}(t)$ were in phase. By doing this, the influence of time delays is eliminated in the comparison of results. Here, Figure 13 presents the average SegSDR over all frequencies for each method. In $\mathrm{Prop}_{\mathrm{rem}}$ and $\mathrm{Prop}_{\mathrm{wei}}$, the accuracy of sound extraction varies depending on α. Therefore, in this experiment, we used the maximum performance values of each method ($\mathrm{Prop}_{\mathrm{rem}}$ or $\mathrm{Prop}_{\mathrm{wei}}$).



(a) Printed pattern A (b) Printed pattern B (c) Printed pattern C

(d) Printed pattern D (e) Printed pattern E (f) Printed pattern F

Figure 13: The mean SegSDR over all frequencies.

We compared the sound quality of the extracted sound in two conditions: when white noise was radiated from speaker 2 and when it was not, as shown in Tables 4 and 5. The results presented in Tables 4 and 5 confirm that there is

no difference in the quality of the extracted sound, regardless of the presence of noise sources. This demonstrates that the sound extracted by the visual microphone contains minimal noise, provided there are no noise sources near the camera. The following discusses the experimental results conducted under the condition without a noise source.

Table 4: Comparison of sound extraction accuracy in terms of SegSDR with printed pattern A (under condition with white noise).

Condition		Frequency [Hz]						
		300	500	700	900	1,100	1,300	1,500
Conv		7.24	3.76	1.73	0.17	**-2.76**	-2.78	-2.88
Prop_{wei}		8.30	3.69	1.04	-0.51	-2.77	-2.85	-2.88
Prop_{rem}	$\alpha = 0.5$	7.25	3.69	1.80	0.20	-2.78	-2.77	-2.91
	$\alpha = 1.0$	7.21	3.70	1.71	0.20	-2.79	**-2.76**	-2.93
	$\alpha = 1.5$	7.46	4.10	1.99	0.39	-2.77	-2.78	-2.89
	$\alpha = 2.0$	8.16	4.44	**2.13**	0.46	-2.81	-2.81	-2.87
Prop_{com}	$\alpha = 0.5$	8.26	3.69	1.07	-0.48	-2.79	-2.88	-2.87
	$\alpha = 1.0$	8.27	3.72	1.04	-0.48	-2.79	-2.81	-2.87
	$\alpha = 1.5$	**8.37**	3.92	1.21	-0.32	-2.83	-2.90	-2.86
	$\alpha = 2.0$	7.96	**4.73**	2.06	**0.48**	-2.81	-2.85	**-2.85**

Table 5: Comparison of sound extraction accuracy in terms of SegSDR with printed pattern A.

Condition		Frequency [Hz]						
		300	500	700	900	1,100	1,300	1,500
Conv		6.68	4.28	1.13	0.19	-0.74	-1.97	-2.88
Prop_{wei}		7.97	4.86	0.76	-0.38	-1.43	-2.21	-2.85
Prop_{rem}	$\alpha = 0.5$	6.64	4.28	1.15	0.19	-0.74	-1.97	-2.88
	$\alpha = 1.0$	6.67	4.25	1.11	0.16	-0.76	-1.93	-2.85
	$\alpha = 1.5$	6.80	4.40	1.34	0.29	-0.71	-1.87	-2.83
	$\alpha = 2.0$	7.38	**5.12**	**1.58**	**0.87**	**-0.46**	**-1.65**	-2.62
Prop_{com}	$\alpha = 0.5$	**8.12**	4.86	0.78	-0.38	-1.43	-2.21	-2.85
	$\alpha = 1.0$	8.01	4.88	0.72	-0.37	-1.42	-2.27	-2.85
	$\alpha = 1.5$	8.10	5.05	0.85	-0.29	-1.35	-2.20	-2.84
	$\alpha = 2.0$	7.53	5.71	1.47	0.60	-1.01	-1.79	**-2.58**

Table 6: Comparison of sound extraction accuracy in terms of SegSDR with printed pattern B.

Condition		Frequency [Hz]						
		300	500	700	900	1,100	1,300	1,500
Conv		0.42	0.05	-1.40	-1.76	-2.39	-2.63	-2.92
Prop_{wei}		-0.65	-0.36	-1.39	-1.99	-2.46	-2.61	-3.03
Prop_{rem}	$\alpha = 0.5$	0.37	0.07	-1.42	-1.77	-2.40	-2.62	-2.93
	$\alpha = 1.0$	0.51	0.16	-1.33	-1.71	-2.36	-2.62	-2.93
	$\alpha = 1.5$	1.15	0.75	-1.01	-1.35	-2.34	-2.56	**-2.91**
	$\alpha = 2.0$	**1.36**	**0.76**	-1.04	-1.26	-2.25	-2.49	-2.96
Prop_{com}	$\alpha = 0.5$	-0.68	-0.37	-1.40	-1.97	-2.45	-2.58	-3.00
	$\alpha = 1.0$	-0.50	-0.16	-1.30	-1.79	-2.41	-2.59	-3.10
	$\alpha = 1.5$	0.10	0.53	**-0.77**	-1.35	**-2.21**	-2.44	-3.01
	$\alpha = 2.0$	0.46	0.57	-0.79	**-1.24**	-2.29	**-2.42**	-2.95

Table 7: Comparison of sound extraction accuracy in terms of SegSDR with printed pattern C.

Condition		Frequency [Hz]						
		300	500	700	900	1,100	1,300	1,500
Conv		8.52	5.17	3.92	1.97	-0.19	-0.86	-2.89
Prop_{wei}		**9.32**	6.08	3.85	2.16	-0.35	-1.00	**-2.87**
Prop_{rem}	$\alpha = 0.5$	8.50	5.26	3.93	2.00	-0.17	-0.82	-2.89
	$\alpha = 1.0$	8.46	5.34	3.95	2.02	-0.13	-0.80	-2.90
	$\alpha = 1.5$	8.40	5.39	**3.98**	2.07	**-0.07**	**-0.79**	-2.90
Prop_{com}	$\alpha = 0.5$	9.26	6.10	3.84	2.16	-0.33	-1.00	-2.88
	$\alpha = 1.0$	9.20	6.18	3.86	2.17	-0.31	-0.99	**-2.87**
	$\alpha = 1.5$	8.93	**6.21**	3.88	**2.19**	-0.12	-0.91	**-2.87**

Table 8: Comparison of sound extraction accuracy in terms of SegSDR with printed pattern D.

Condition		Frequency [Hz]						
		300	500	700	900	1,100	1,300	1,500
Conv		0.14	0.82	-1.03	-1.61	-2.36	-2.60	-2.80
Prop_{wei}		**3.51**	1.13	-0.48	-1.70	-2.48	-2.63	-2.86
Prop_{rem}	$\alpha = 0.5$	0.17	0.84	-1.03	-1.60	-2.36	-2.61	-2.78
	$\alpha = 1.0$	-0.00	0.82	-0.97	-1.66	-2.40	-2.61	**-2.77**
	$\alpha = 1.5$	-0.14	**1.25**	**-0.14**	**-1.25**	**-2.21**	**-2.54**	-2.87
Prop_{com}	$\alpha = 0.5$	3.34	1.10	-0.61	-1.69	-2.47	-2.62	-2.84
	$\alpha = 1.0$	3.37	1.09	-0.47	-1.76	-2.55	-2.62	-2.85
	$\alpha = 1.5$	2.52	0.59	-0.44	-1.76	-2.54	-2.61	-2.91

Table 9: Comparison of sound extraction accuracy in terms of SegSDR with printed pattern E.

Condition		Frequency [Hz]						
		300	500	700	900	1,100	1,300	1,500
Conv		6.87	4.91	2.15	0.88	-0.59	-1.56	-2.05
Prop_{wei}		7.96	5.77	2.52	1.11	-0.33	-1.41	-1.99
Prop_{rem}	$\alpha = 0.5$	6.90	4.89	2.16	0.87	-0.59	-1.56	-2.06
	$\alpha = 1.0$	8.43	5.82	2.80	1.39	-0.15	-1.31	-1.87
	$\alpha = 1.5$	8.98	6.27	2.94	1.45	-0.14	-1.16	-1.82
	$\alpha = 2.0$	9.04	6.35	2.93	1.48	-0.16	-1.19	-1.83
	$\alpha = 2.5$	9.26	6.40	2.88	1.46	-0.22	-1.17	-1.86
	$\alpha = 3.0$	9.18	6.36	2.86	1.29	-0.30	-1.24	-1.86
Prop_{com}	$\alpha = 0.5$	7.95	5.79	2.57	1.11	-0.37	-1.48	-1.97
	$\alpha = 1.0$	9.01	6.72	3.32	1.67	**0.07**	-1.16	-1.77
	$\alpha = 1.5$	9.30	7.05	**3.39**	**1.74**	0.00	-1.01	**-1.71**
	$\alpha = 2.0$	9.22	**7.11**	3.34	1.69	0.01	-1.02	-1.74
	$\alpha = 2.5$	9.30	7.04	3.30	1.70	0.00	**-1.01**	-1.73
	$\alpha = 3.0$	**9.31**	6.93	3.20	1.47	-0.11	-1.11	-1.79

The quality of extracted sound was compared based on time waveforms in Figures 7–12. In all printed patterns, Prop_{rem} improves SDR by removing out-of-focus areas where displacement is difficult to measure accurately. This demonstrates that displacement can be accurately measured in the focused areas calculated by Prop_{rem}, confirming the effectiveness of Prop_{rem}. It was observed that Prop_{wei} improved SDR in printed patterns A, B, C, E, and

Table 10: Comparison of sound extraction accuracy in terms of SegSDR with printed pattern F.

Condition		Frequency [Hz]						
		300	500	700	900	1,100	1,300	1,500
Conv		4.29	2.81	0.83	-0.33	-1.53	-1.96	-2.12
Prop_{wei}		5.18	3.24	0.73	-0.26	-1.57	-1.92	-2.10
Prop_{rem}	$\alpha = 0.5$	4.29	2.84	0.81	-0.34	-1.55	-1.99	-2.08
	$\alpha = 1.0$	5.75	3.80	1.44	0.12	-1.34	-1.76	-1.95
	$\alpha = 1.5$	6.56	4.33	1.55	0.38	-1.20	-1.65	-1.86
	$\alpha = 2.0$	6.43	4.29	1.54	0.25	-1.28	-1.72	-1.91
	$\alpha = 2.5$	5.75	3.63	0.91	-0.13	-1.53	-1.85	-2.05
Prop_{com}	$\alpha = 0.5$	5.15	3.24	0.77	-0.20	-1.55	-1.99	-2.06
	$\alpha = 1.0$	6.36	4.10	1.43	0.18	-1.35	-1.69	-1.95
	$\alpha = 1.5$	6.96	4.57	**1.62**	**0.37**	**-1.08**	**-1.57**	**-1.71**
	$\alpha = 2.0$	**6.97**	**4.62**	**1.62**	0.32	-1.17	-1.59	-1.82
	$\alpha = 2.5$	6.20	3.94	0.90	-0.04	-1.43	-1.75	-1.97

F. Most of these patterns feature prominent vertical edges or complex designs. Prop_{wei} emphasizes displacement at the edges of the captured video, which explains the confirmed effectiveness of Prop_{wei} when vertical edges are pronounced or complex patterns are used. The effectiveness of Prop_{com} was particularly confirmed for printed patterns E and F. Figures 11 (a) and (f) show that the time waveform in (f) exhibits reduced distortion compared to that in (a), with a similar trend observed in Figures 12 (a) and (f).

Tables 5–10 show the evaluation results of the quality of the extracted sound in terms of SegSDR for each printed pattern. We compare the sound quality of the extracted sound by Prop_{rem} and Conv. At most frequencies, the SegSDR of Prop_{rem} was higher than that of Conv, as shown in Table 5–10. This indicates that Prop_{rem} demonstrates consistent performance regardless of the printed pattern. However, at 1,300 and 1,500 Hz, the effect of Prop_{rem} is slight. This is likely due to the difficulty in accurately measuring displacement of high-frequency components, as mentioned in the introduction. It can be considered that the vibration is too small, making it difficult to accurately measure displacement even in the focused area. Then, we focus on the evaluation results shown in Table 9, 10. A comparison between Conv and Prop_{rem} shows that Prop_{rem} results in a higher SegSDR at most frequencies in most values of α. However, depending on the value of α, the SegSDR for the Prop_{rem} may sometimes be lower than that of Conv. Here, we experimented to demonstrate the extent to which reduced number of rows of captured video affects the smoothing of phase variations. Figure 14 shows how the displacement calculated from a row changes when the number of columns in a captured video of printed pattern E is reduced. In Figure 14, it can be seen that when the number of columns is significantly reduced (i.e., α changes from 0.5 to 1.0), the amplitude value also changes significantly. In the removal method, if α is set too large (especially when $\alpha = 3.0$), the amplitude also changes

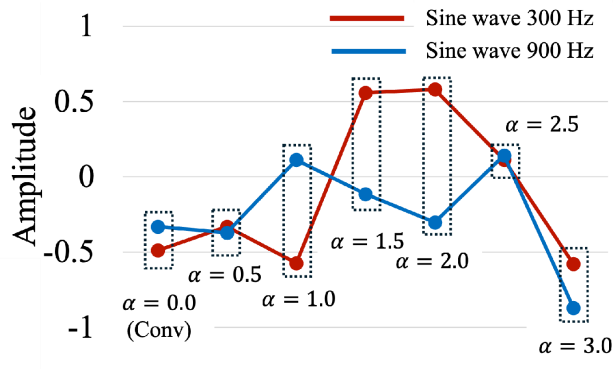


Figure 14: Amplitude for each α.

significantly in Figure 14. Additionally, in the removal method, increasing the α too much leads to a decrease in SegSDR in Tables 9 and 10. This is likely because removing out-of-focused area reduces the number of rows, which weakens the smoothing effect on phase changes, thus Increasing the effect of outliers. Therefore, as mentioned earlier, it was confirmed in this experiment that the reduction in the number of rows increases the influence of outliers, which is the primary cause of the degradation in extracted sound quality.

Prop_{wei} with weighted phase variation applied exhibits lower SegSDR than Conv at almost all frequencies in Tables 6. However, for printed patterns C, E, and F, the SegSDR of Prop_{wei} is higher than that of Conv, further confirming its effectiveness for patterns with vertical features or complex designs. A comparison between Conv and Prop_{com} shows that Prop_{com} results in a higher SegSDR, particularly for printed patterns where the effectiveness of Prop_{wei} was confirmed. From this result, it can be inferred that Prop_{com} is strongly influenced by the effects of Prop_{wei}. When vertical edges are pronounced or when the printed pattern is complex, the SegSDR of Prop_{com} improves significantly compared to that of Conv. This improvement is considered to be due to the combination of weighting based on focal rate and the elimination of out-of-focus areas, which enhanced the contribution of the focused area in sound extraction, thereby enabling more accurate displacement measurement.

Figure 13 shows the evaluation of the quality of the extracted sound in terms of the average of SegSDR over all frequencies. From Figure 13, it can be observed that in most cases, Prop_{com} achieves the highest values. This demonstrates that combining the removal of the out-of-focused area with the measurement of displacement at the edges of the captured video effectively improves the quality of the extracted sound. However, for certain printed patterns, the effectiveness could not be confirmed. Additionally, in Prop_{rem} and

Prop_{com}, the accuracy of sound extraction varies depending on α. Therefore, in this experiment, we used the maximum performance values of each method (Prop_{rem} or Prop_{com}). In the future, it will be necessary to devise a method that can adaptively select α for the sound source.

In summary, we confirmed the effectiveness of Prop_{rem} across most patterns and frequencies, demonstrating that it is a versatile method. On the other hand, the effectiveness of Prop_{wei} and Prop_{com} was only observed under specific conditions, namely when the A4 paper being vibrated by sound printed vertical edges or complex patterns.

4.3 Time complexity

Table 3 shows that the number of columns of captured videos decreases as α increases. By reducing the number of columns, the number of convolutions between the captured images and the Gabor filter in the speech extraction process is reduced, and thus the execution time can be reduced. Here, we compare the processing time between Conv and Prop_{com} in printed patten E. The results are shown in Figure 15. Specifically, when using MATLAB on a computer equipped with an Intel(R) Core(TM) i9-11950H @2.60GHz and 32 GB of RAM, Conv requires approximately 104 seconds to extract sound from about 5 seconds video. On the other hand, Prop_{com} with $\alpha = 3.0$ reduces the processing time to approximately 19 seconds under the same conditions. These results indicate that the method of out-of-focused area removal reduces the execution time without degrading the quality of the extracted sound. We confirmed similar trends with other printed patterns.

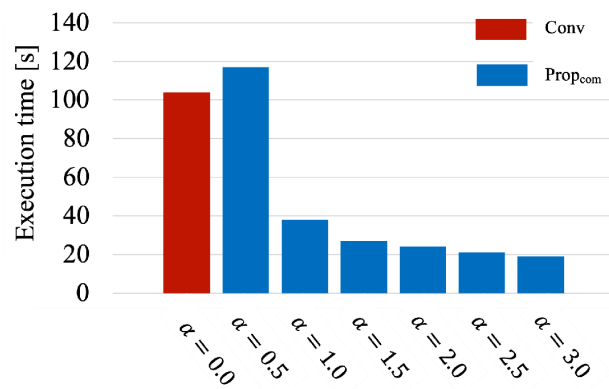


Figure 15: Comparison of execution time.

5 Conclusion

In this paper, sound extraction methods are proposed that emphasize the displacements measured in the focused area of the captured video based on the focal rate. Specifically, there are three methods: out-of-focused area removal, weighted phase variation, and a method combining these two methods. The visual microphone extracts sound by measuring the displacement from video captured of the object being vibrated by sound. In the captured video, there may be area that arise out-of-focused due to the depth of field. Although out-of-focused area may cause degradation in the quality of the extracted sound, this is not taken into account in conventional methods. Therefore, we propose methods to extract sound by emphasis on the focused area where displacement can be measured more accurately than in the out-of-focused area.

Experimental results demonstrate that the proposed methods are effective in improving the sound quality of the visual microphone under certain conditions. In particular, it can be observed that the sound was accurately extracted with small distortions in the complex pattern to print on A4 paper.

In the future, we will conduct experiments to confirm the accuracy of the sound quality improvement of the extracted sound by using sound sources with multiple frequency components.

Acknowledgments

This work was partly supported by Ritsumeikan University R-GIRO and RARA, and JSPS KAKENHI Grant Numbers JP21H03488, JP23H03425, JP23K21691, JP23K28115, and JP24K20803.

References

[1] O. Ait-Aider, N. Andreff, J. M. Lavest, and P. Martinet, "Exploiting Rolling Shutter Distortions for Simultaneous Object Pose and Velocity Computation Using a Single View", in *Proceedings of Fourth IEEE International Conference on Computer Vision Systems (ICVS'06)*, 2006, 35–41, DOI: 10.1109/ICVS.2006.25.

[2] S. Baker, E. Bennett, S. B. Kang, and R. Szeliski, "Removing rolling shutter wobble", in *Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, 2392–9, DOI: 10.1109/CVPR.2010.5539932.

[3] Y. Dai, H. Li, and L. Kneip, "Rolling Shutter Camera Relative Pose: Generalized Epipolar Geometry", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 4132–40, DOI: 10.1109/CVPR.2016.448.

[4] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The Visual Microphone: Passive Recovery of Sound from Video", *ACM Transactions on Graphics*, 33(4), 2014, 1–10, ISSN: 0730-0301.

[5] H. Foroosh, J. B. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration", *IEEE Transactions on Image Processing*, 11(3), 2002, 188–200.

[6] T. Gautama and M. A. Van Hulle, "A phase-based approach to the estimation of the optical flow field using spatial filtering", *IEEE Transactions on Neural Networks*, 13(5), 2002, 1127–36, DOI: 10.1109/TNN.2002.1031944.

[7] M. Grundmann, V. Kwatra, D. Castro, and I. Essa, "Calibration-free rolling shutter removal", in *Proceedings of 2012 IEEE International Conference on Computational Photography (ICCP)*, 2012, 1–8, DOI: 10.1109/ICCPhot.2012.6215213.

[8] A. Janssen, R. Veldhuis, and L. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(2), 1986, 317–30, DOI: 10.1109/TASSP.1986.1164824.

[9] Multimedia Signal Processing and Security Lab, University of Salzburg, Salzburg, Austria, "Salzburg Texture Image Database (STex)", https://wavelab.at/sources/STex/, 2023.

[10] H. Nakano, Y. Geng, K. Iwai, and T. Nishiura, "Sound Quality Improvement in Visual Microphone by Emphasizing Focused Area Based on Focal Rate", in *Proceedings of 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024 (in press).

[11] H. Nakano, T. Yoshizawa, Y. Geng, K. Iwai, and T. Nishiura, "Speech Quality Improvement Utilizing Out-of-Focus Areas in Rolling-Shutter Video on Speech Extraction", in *Proceedings of 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, 2320–5, DOI: 10.1109/APSIPAASC58517.2023.10317150.

[12] J. Portilla and E. P. Simoncelli, "A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients", *International Journal of Computer Vision*, 40, 2000, 49–70, DOI: 10.1023/A:1026553619983.

[13] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms", *IEEE Transactions on Information Theory*, 38(2), 1992, 587–607, DOI: 10.1109/18.119725.

[14] K. Terano, H. Shindo, K. Iwai, T. Fukumori, and T. Nishiura, "Sound capture from rolling-shuttered visual camera based on edge detection", in *Proceedings of 23rd International Congress on Acoustics*, 2019, 2878–84.

[15] O. R. Vincent, O. Folorunso, *et al.*, "A descriptive algorithm for sobel image edge detection", in *Proceedings of Information Science & IT Education Conference (InSITE)*, Vol. 40, 2009, 97–107.

[16] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, "Phase-Based Video Motion Processing", *ACM Transactions on Graphics*, 32(4), 2013, 1–10.

[17] N. Wadhwa, H.-Y. Wu, A. Davis, M. Rubinstein, E. Shih, G. J. Mysore, J. G. Chen, O. Buyukozturk, J. V. Guttag, W. T. Freeman, *et al.*, "Eulerian video magnification and analysis", *Communications of the ACM*, 60(1), 2016, 87–95.

[18] A. Yoshida, H. Shindo, K. Terano, T. Fukumori, and T. Nishiura, "Interpolation of acoustic signals in sound capture with rolling-shuttered visual camera", in *Forum Acusticum*, 2020, 39–45.

[19] T. Yoshizawa, A. Yoshida, K. Iwai, and T. Nishiura, "Speech extraction with RGB-intensity gradient on rolling-shutter video", in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Vol. 263, No. 5, Institute of Noise Control Engineering, 2021, 1095–106.