Sound Quality Improvement in Visual Microphone by Emphasizing Focused Area Based on Focal Rate

Hayata Nakano*, Yuting Geng*, Kenta Iwai* and Takanobu Nishiura*
* Ritsumeikan University, Osaka, Japan
E-mail: {is0516vr@ed, geng@fc, iwai18sp@fc, nishiura@is}.ritsumei.ac.jp

***Abstract*—This paper presents a method of improving the sound quality of visual microphone by emphasizing the in-focus area in a captured video. Visual microphone extracts sound by measuring the displacement of a vibrating object from the video of an object vibrating due to sound. In the captured video, there may arise blurred (out-of-focus) areas may arise due to depth of field. In these out-of-focus areas of the captured video, it is difficult to accurately measure the displacement of the vibrating object, which may reduce the sound quality of the extracted sound. However, out-of-focus areas are not taken into account in conventional sound extraction methods. We propose a method to improve the sound quality of the extracted sound in the visual microphone by emphasizing the in-focus area in the captured video. The proposed method emphasizes the measured displacement in the in-focus area of the captured video by using the focal rate that represents the degree of focus. Experimental results show that the proposed method improves the quality of the extracted sound compared to the conventional methods.**

I. 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 noise around itself in addition to the target sound. The visual microphone [1], [2] was proposed to address this problem. The visual microphone extracts a sound signal from a video of the object vibrating due to the target sound captured by a consumer camera. The sound signal extracted by the visual microphone contains little noise that occurs near the camera.

Many consumer-level cameras are equipped with rolling shutter image sensors [3]. These cameras expose and read pixels from the top row of an image to the bottom. Therefore, when capturing an object vibrating due to sound, rolling shutter distortion [3] arises in the captured images as a result of differences in exposure start times. 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-focus (i.e., blurred) areas arise in the captured video. Therefore, when extracting sound including out-of-focus areas in the captured video, it is difficult to accurately measure the displacement of the vibrating object, leading to degradation in the quality of the extracted sound signal. However, out-of-focus areas are not taken into account in conventional sound extraction methods [1], [2].

Taking into account the out-of-focus areas in the captured video, this paper aims to improve the sound quality of the extracted sound in the visual microphone by emphasizing the in-focus area in the captured video. In our previous research [4], we proposed two methods: removing out-of-focus areas and weighting phase variation. By improving the sound quality of the extracted sound using our previous methods [4], we further emphasize the measured displacement in the in-focus area based on focal rate, which represents the degree of focus, to improve the sound quality of the extracted sound in the visual microphone. We conducted evaluation experiments to verify the effectiveness of the proposed method for sound extraction.

II. Conventional sound extraction method in visual microphone

The conventional method for sound extraction in the visual microphone [1] is illustrated in Fig. 1. The conventional method extracts sound signal $u(t)$ by calculating the displacement of the object using a complex steerable pyramid [5] on a video $I(x, y, n)$ of the object vibrating due to sound, where x and y are the row and column indices, n is the frame index, and t is the time index. Here, the vibration of the object appears as phase variation [6], [7], [8] enabling us to extract sound signal based on the phase variation of each row in each frame. The displacement of the object is obtained based on the phase variation calculated using the complex steerable pyramid.

The complex steerable pyramid consists of sub-band images for each scale r and orientation θ. Amplitude $A(r, \theta, x, y, n)$ and phase $\phi(r, \theta, x, y, n)$ are calculated by applying the complex steerable pyramid to the video $I(x, y, n)$. $A(r, \theta, x, y, n)$ and $\phi(r, \theta, x, y, n)$ are downsampled to 2^{-r} times the original resolution per scale r. As the conventional method uses the phase variation of each row in each video frame, the orientation θ is fixed to 0. Therefore, $A(r, \theta, x, y, n)$ and $\phi(r, \theta, x, y, n)$ are replaced by $A(r, x, y, n)$ and $\phi(r, x, y, n)$.

The phase variation $\phi_{\mathrm{v}}(r, x, y, n)$ is calculated from the phase difference with the reference frame $\phi(r, x, y, n_0)$, where n_0 denotes the index representing the reference frame. We calculate the row-wise weighted average of $\phi_{\mathrm{v}}(r, x, y, n)$ to obtain $\Phi(r, y, n)$. 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.

The frame gaps arise due to the characteristics of rolling shutter image sensors [3]. Therefore, an autoregressive model [9] is utilized to interpolate the frame gaps in the extracted signal $\tilde{\Phi}(r, t)$. Finally, the extracted sound signal $u(t)$ is

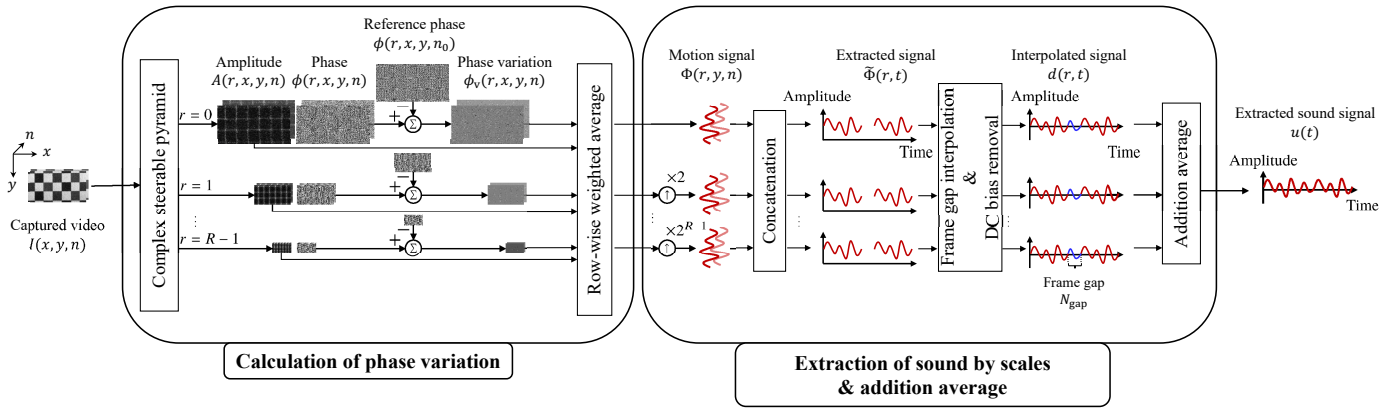


Fig. 1. Overview of conventional sound extraction method of visual microphone [1].

obtained by calculating the average of interpolated signal $d(r,t)$ at each scale r.

Out-of-focus areas are not taken into account in conventional methods [1], [2]. Hence, it is difficult to accurately calculate the phase variation when there are out-of-focus areas in the captured video. This causes the sound quality of the extracted sound to deteriorate. In this paper, we propose to improve the sound quality of the extracted sound in the visual microphone by emphasizing the in-focus area in the captured video.

III. Proposed Sound Extraction Method for Visual Microphone

Our proposed sound extraction method emphasizes the in-focus area in the captured video to improve the sound quality of the extracted sound. The proposed method is based on a combination of two methods, removing out-of-focus areas (Removed) and weighting phase variation (Weighted), proposed in [4]. The procedure of the proposed method is illustrated in Fig. 2.

Step 1: Calculation of Focal Rate and Out-of-focus Areas Removal

As the phase variation is not measured accurately in the out-of-focus areas of the captured video, we only utilize the in-focus area to extract the sound. Here, in the out-of-focus areas, the edges of the captured video are blurred, whereas the edges are relatively clear in the in-focus area. Therefore, the focal rate is calculated using a Sobel filter, which is commonly used as an edge enhancement filter [10]. This method only uses the displacement in the horizontal direction; thus, a horizontal Sobel filter is used. In out-of-focus areas, the change in pixel values is smooth, resulting in small 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 is calculated as

$$M(x,y,n) = |K(x,y) * I(x,y,n)|, \tag{1}$$

where $*$ denotes convolution operator and $K(x,y)$ denotes a 3×3 horizontal Sobel filter.

In this paper, we assume that the in-focus area does not change significantly between frames of the captured video. Therefore, we remove the out-of-focus areas by using the focal rate $M(x,y,n_0)$ calculated from the n_0-th frame.

First, the column-wise mean of $M(x,y,n_0)$ is calculated as

$$\overline{M}(x,n_0) = \frac{1}{H}\sum_{y=0}^{H-1} |M(x,y,n_0)|, \tag{2}$$

where H denotes the number of rows in $I(x,y,n)$. Next, we define $\boldsymbol{F}$ as the set of column indices x such that $\overline{M}(x,n_0)$ is larger than the threshold value $\overline{M}_{\mathrm{th}}$:

$$\boldsymbol{F} := \{x \mid \overline{M}(x,n_0) \geq \overline{M}_{\mathrm{th}}\}, \tag{3}$$

$$\overline{M}_{\mathrm{th}} = \frac{\alpha}{W}\sum_{x=0}^{W-1} \overline{M}(x,n_0), \tag{4}$$

where $\overline{M}_{\mathrm{th}}$ is calculated as α times the row-wise means $\overline{M}(x,n_0)$. The effect of removing out-of-focus areas can be amplified by increasing this α. The column indices indicating the start point F_{start} and endpoint F_{end} of the in-focus area are calculated as

$$(F_{\mathrm{start}}, F_{\mathrm{end}}) = \left(\min_{x\in\boldsymbol{F}} x, \max_{x\in\boldsymbol{F}} x\right). \tag{5}$$

Finally, we remove the out-of-focus areas from the captured video $I(x,y,n)$ and calculate $I_{\mathrm{focal}}(\tilde{x},y,n)$ which is the video of the in-focus area as

$$I_{\mathrm{focal}}(\tilde{x},y,n) = I(\tilde{x}+F_{\mathrm{start}},y,n), 0 \leq \tilde{x} \leq \tilde{W}-1, \tag{6}$$

$$\tilde{W} = F_{\mathrm{end}} - F_{\mathrm{start}} + 1, \tag{7}$$

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

Step 2: Calculation of Phase Variation

As in the conventional method [1], amplitude $A(r,\tilde{x},y,n)$ and phase $\phi(r,\tilde{x},y,n)$ are calculated by applying the complex steerable pyramid to $I_{\mathrm{focal}}(\tilde{x},y,n)$. $A(r,\tilde{x},y,n)$ and

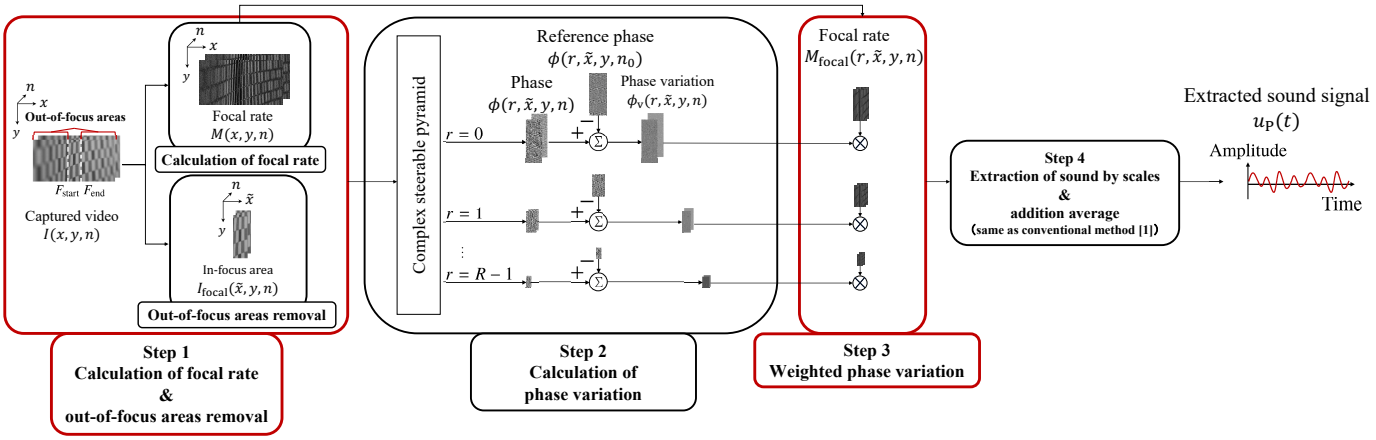


Fig. 2. Overview of proposed sound extraction method for visual microphone.

$\phi(r,\tilde{x},y,n)$ are downsampled to 2^{-r} times the original resolution per scale r. Then, the numbers of rows and columns in $A(r,\tilde{x},y,n)$ and $\phi(r,\tilde{x},y,n)$ for each scale r are given by

$$\tilde{W}_r = 2^{-r}\tilde{W}, \tag{8}$$

$$H_r = 2^{-r}H, \tag{9}$$

where $\tilde{W}$ and H denote the numbers of columns and rows in $I_{\mathrm{focal}}(\tilde{x},y,n)$. The phase variation $\phi_{\mathrm{v}}(r,\tilde{x},y,n)$ is calculated from the phase difference with the reference frame $\phi(r,\tilde{x},y,n_0)$ as

$$\phi_{\mathrm{v}}(r,\tilde{x},y,n) = \phi(r,\tilde{x},y,n) - \phi(r,\tilde{x},y,n_0). \tag{10}$$

Step 3: Weighted Phase Variation

The phase variation $\phi_{\mathrm{v}}(r,\tilde{x},y,n)$ obtained in Step 2 is weighted by the focal rate $M(x,y,n)$. The focal rate $M_{\mathrm{focal}}(\tilde{x},y,n)$ of the in-focus area is calculated as

$$M_{\mathrm{focal}}(\tilde{x},y,n) = M(\tilde{x}+F_{\mathrm{start}},y,n),\ 0 \le \tilde{x} \le \tilde{W}-1. \tag{11}$$

The focal rate $M_{\mathrm{focal}}(r,\tilde{x},y,n)$ for each scale r is calculated. 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. In the conventional method, the phase variation $\phi_{\mathrm{v}}(r,x,y,n)$ is weighted by the square of the amplitude $A(r,x,y,n)$ since the phase can be accurately estimated for pixels with a large amplitude. The row-wise weighted average signal (the motion signal) $\Phi(r,y,n)$ is calculated as

$$\Phi(r,y,n) = \sum_{x=0}^{W_r-1} A^2(r,x,y,n)\phi_{\mathrm{v}}(r,x,y,n), \tag{12}$$

where W_r denotes the number of columns in $I(x,y,n)$ for each scale r.

In the proposed method, (12) is replaced by (13), weighting $\phi_{\mathrm{v}}(r,\tilde{x},y,n)$ by $M_{\mathrm{focal}}(r,\tilde{x},y,n)$ to emphasize the phase variation at the edges in $I_{\mathrm{focal}}(\tilde{x},y,n)$.

$$\Phi_{\mathrm{P}}(r,y,n) = \frac{1}{\tilde{W}_r}\sum_{\tilde{x}=0}^{\tilde{W}_r-1} M^2_{\mathrm{focal}}(r,\tilde{x},y,n)\phi_{\mathrm{v}}(r,\tilde{x},y,n). \tag{13}$$

As the length of $\Phi_{\mathrm{P}}(r,y,n)$ 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, $\Phi_{\mathrm{P}}(r,y,n)$ for each r is a two-dimensional signal. Therefore, the averaged phase variation is transformed to a one-dimensional signal $\tilde{\Phi}_{\mathrm{P}}(r,t)$ for each r as

$$\tilde{\Phi}_{\mathrm{P}}(r,t) = \Phi_{\mathrm{P}}(r,y,n), \tag{14}$$

$$t = y + (H + N_{\mathrm{gap}})n, \tag{15}$$

where N_{gap} is the number of samples in the frame gap. As noted in Section 2, the frame gaps arise due to the characteristics of rolling shutter image sensors [3]. Therefore, an autoregressive model [9] is utilized to interpolate the frame gaps in the extracted signal $\tilde{\Phi}(r,t)$ to obtain an interpolated signal $d_{\mathrm{P}}(r,t)$.

Step 4: Extraction of Sound by Scales

The extracted sound signal $u_{\mathrm{P}}(t)$ is obtained by calculating the average of interpolated signal $d_{\mathrm{P}}(r,t)$ at each scale via

$$u_{\mathrm{P}}(t) = \frac{1}{R}\sum_{r=0}^{R-1} d_{\mathrm{P}}(r,t), \tag{16}$$

where R denotes the scale number of the complex steerable pyramid.

IV. Evaluation Experiments

We conducted evaluation experiments to investigate the effectiveness of the proposed method by comparing the sound extraction using the proposed method (Prop.) with the conventional methods (Conv. [1], Removed [4], and Weighted [4]).

A. Experimental Conditions for Sound Extraction

Figs. 3 shows the experimental setup and equipment arrangement, and 4 shows an A4 paper with a printed pattern used as the vibrating object. Moreover, Tables I, II and III show the experimental conditions, equipment and number of rows of video in each α in (4). As shown in Fig. 3, the A4 paper placed in front of a loudspeaker was captured with a rolling shutter camera, and the sound signal was extracted from the

captured video. To obtain the video with out-of-focus areas, the A4 paper was captured at 30°. This takes into account that capturing the subject at this angle causes out-of-focus areas in the video resulting from the relationship between focal length and depth of field. The pattern shown in Fig. 4 was printed on the A4 paper from the Salzburg Texture Image Database (STex) [11]. 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. Fig. 5 shows one frame of the captured video.

We used sine waves as the sound source. Additionally, each sine wave was radiated from a loudspeaker for 5 s. In this experiment, we set $n_0 = 0$ as the reference frame and $r_0 = 0$ as a reference scale.

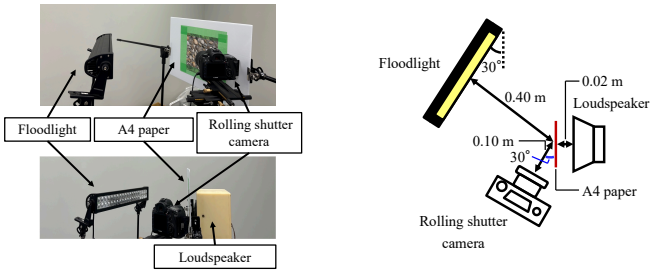


Fig. 3. Experimental setup and arrangement of experimental equipment.

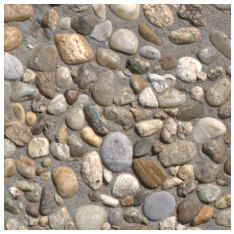


Fig. 4. Printed pattern.

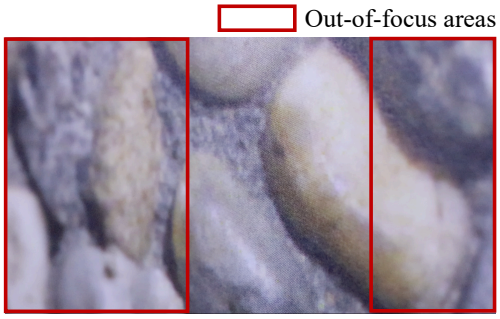


Fig. 5. One frame of captured video.

TABLE I
EXPERIMENTAL CONDITION

Ambient noise level	L_A=41.8 dB
Temperature / Humidity	20.8°C/51.7%
Sound source	Sine wave (300, 500, ⋯, 1,500 Hz)
Sound pressure level	85 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 II
EXPERIMENTAL EQUIPMENT

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

B. Experimental Results of Sound Extraction

We evaluated the sound quality of extracted sound in terms of signal-to-distortion ratio (SDR). The SDR is calculated as

$$\mathrm{SDR} = 10\log 10 \left[\frac{\sum_{t=0}^{T-1} u_{\mathrm{c}}^2(t)}{\sum_{t=0}^{T-1} \{u_{\mathrm{c}}(t) - \lambda u(t)\}^2} \right], \quad (17)$$

$$\lambda = \sqrt{\frac{\sum_{t=0}^{T-1} u_{\mathrm{c}}^2(t)}{\sum_{t=0}^{T-1} u^2(t)}}, \quad (18)$$

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

Fig. 6 presents the evaluation results of the quality of the extracted sound in terms of SDR. The values in Fig. 6 indicate the average SDR. Fifty signals are randomly selected from the sound signal extracted by each method and the average SDR of the obtained Fifty signals is calculated.

Fig. 6 shows that the average SDR of Prop. is the highest at 500 – 1,500 Hz. This can be explained by the weighting based on the focal rate and the removal of out-of-focus areas, which can increase the contribution rate of the in-focus area in sound extraction, thereby enabling accurate measurement of the displacement. In particular, the accuracy of the sound quality improvement of the extracted sound of Prop. is highest at 500 Hz. Meanwhile, the average SDR of Conv. is highest at 300 Hz. However, it should be noted that the average SDR at 300 Hz is considerably lower than at other frequencies for all methods. Therefore, these results demonstrate that emphasizing the in-focus area is effective in improving the sound quality of the extracted sound, provided the sound quality is maintained at a certain level.

Fig. 7 presents the time waveforms of the 500 Hz sine waves extracted by the Conv., Weighted, Removed, and Prop. methods. Specifically, each figure shows the time waveform of 0.01 s signals from the extracted sound and the sound source, depicted as red lines and blue dotted lines respectively. The threshold values α for the Removed and Prop. methods are

Condition	Frequency [Hz] 300	500	700	900	1,100	1,300	1,500
Conv.	**-1.39**	**7.08**	**3.82**	**3.51**	**2.04**	**1.42**	**1.97**
Weighted	**-2.22**	**7.25**	**4.02**	**3.84**	**2.36**	**1.77**	**2.20**
Removed ($\alpha = 0.5$)	**-1.44**	**7.03**	**3.89**	**3.53**	**2.05**	**1.50**	**2.15**
Removed ($\alpha = 1.0$)	**-1.46**	**7.75**	**4.03**	**3.93**	**2.38**	**1.66**	**2.51**
Removed ($\alpha = 1.5$)	**-1.51**	**7.81**	**4.07**	**3.96**	**2.50**	**1.68**	**2.31**
Removed ($\alpha = 2.0$)	**-1.56**	**7.88**	**4.01**	**3.86**	**2.48**	**1.69**	**2.25**
Removed ($\alpha = 2.5$)	**-1.6**	**8.12**	**4.08**	**3.86**	**2.39**	**1.67**	**2.17**
Removed ($\alpha = 3.0$)	**-1.65**	**8.20**	**3.92**	**3.81**	**2.16**	**1.57**	**2.06**
Removed ($\alpha = 3.5$)	**-1.97**	**6.81**	**3.33**	**3.05**	**1.94**	**1.66**	**1.98**
Prop. ($\alpha = 0.5$)	**-2.22**	**7.25**	**4.02**	**3.84**	**2.36**	**1.77**	**2.2**
Prop. ($\alpha = 1.0$)	**-2.24**	**8.57**	**4.12**	**3.96**	**2.48**	**1.85**	**2.32**
Prop. ($\alpha = 1.5$)	**-2.29**	**8.85**	**4.21**	**4.40**	**2.82**	**2.01**	**2.75**
Prop. ($\alpha = 2.0$)	**-2.31**	**8.88**	**4.25**	**4.33**	**2.61**	**2.04**	**2.64**
Prop. ($\alpha = 2.5$)	**-2.34**	**8.98**	**4.28**	**4.27**	**2.73**	**2.00**	**2.67**
Prop. ($\alpha = 3.0$)	**-2.36**	**9.09**	**4.27**	**4.27**	**2.69**	**2.00**	**2.57**
Prop. ($\alpha = 3.5$)	**-2.36**	**9.22**	**4.16**	**4.24**	**2.49**	**1.90**	**2.54**

Fig. 6. Comparison of sound extraction accuracy in terms of signal-to-distortion ratio (SDR).

TABLE III
NUMBER OF ROWS OF VIDEO IN EACH α

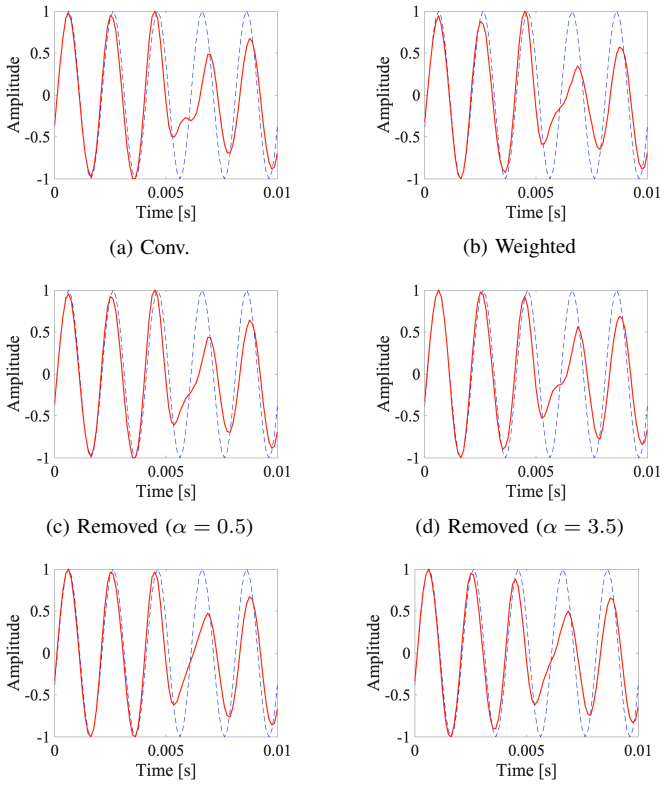
α in Eq. (4)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
In-focus area W	1,871	1,046	292	254	220	192	160

set to $\alpha = 0.5, 3.5$ in (4). The results show that the highest accuracy of sound extraction is achieved using Prop. method. This is because the value used to weight the phase variation was changed from amplitude to focal rate, which enabled the weighting of pixels with a higher degree of confidence.

These results show that the proposed method improves the sound quality of the extracted sound for sine waves of several frequencies.

V. CONCLUSION

In this paper, we have presented that the sound quality of the visual microphone can be improved by emphasizing the in-focus area of the captured video. Utilizing the focal rate that represents the degree of focus, we proposed a sound extraction method that emphasizes the displacement measured in the in-focus area of the captured video. Experimental results demonstrate that the proposed method is effective for improving the sound quality of the visual microphone. In the future, we will conduct experiments under different conditions, changing the pattern and material of the vibrating object, as well as the distance between the vibrating object and the sound source.



(a) Conv. (b) Weighted

(c) Removed ($\alpha = 0.5$) (d) Removed ($\alpha = 3.5$)

(e) Prop. ($\alpha = 0.5$) (f) Prop. ($\alpha = 3.5$)

Fig. 7. Time waveform (500 Hz) of sound extracted by each method (red line: extracted sound, blue dotted line: sound source) .

ACKNOWLEDGMENT

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] 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*, vol. 33, no. 4, pp. 1–10, 2014.

[2] 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*, pp. 2878–2884, 2019.

[3] 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)*, pp. 35–41, 2006.

[4] 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)*, pp. 2320–2325, 2023.

[5] J. Portilla and E. P. Simoncelli, "A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients," *International Journal of Computer Vision*, vol. 40, pp. 49–70, 2000.

[6] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, "Phase-Based Video Motion Processing," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 1–10, 2013.

[7] 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*, vol. 13, no. 5, pp. 1127–1136, 2002.

[8] H. Foroosh, J. B. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration," *IEEE Transactions on Image Processing*, vol. 11, no. 3, pp. 188–200, 2002.

[9] 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*, vol. ASSP-34, no. 2, pp. 317–330, 1986.

[10] 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, pp. 97–107, 2009.

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