Sound Recovery Considering the Vibration Direction of an Object in a Video

Yohei Fuse
Graduate School of System Informatics, Kobe University
Japan
y.fuse@stu.kobe-u.ac.jp

Yusuke Yasumi
Graduate School of System Informatics, Kobe University
Japan
yasumi@me.cs.scitec.kobe-u.ac.jp

Tetsuya Takiguchi
Graduate School of System Informatics, Kobe University
Japan
takigu@kobe-u.ac.jp

***Abstract*—When a sound hits an object, it causes the surface of that object to vibrate. Some research has been carried out on the recovering of sounds by extracting the vibrations that have been recorded on high-speed videos. This research is expected to be applied in the field of surveillance and security because sounds can be recorded from far away. The vibration of objects due to sound is so fast and minute that it is invisible, but it is possible to observe the changes in objects as the movement of each pixel by using the high-speed video. In this paper, we propose a sound-recovery method focusing on the vibration direction of the object. It is considered that a better sound can be obtained by trying to recover the sound based on the direction of the largest vibration. First, the sound is recovered in a certain direction (initial direction). Next, the sound is recovered in the vibration direction that has the largest correlation with the pre-first recovered sound. Repeating this process, the sound can be recovered in the direction of the largest vibration. We recovered sounds from several objects in videos and ascertained the effectiveness of the method.**

***Index Terms*—sound recovery, high-speed video, vibration direction**

I. Introduction

When a sound hits an object, it causes the surface of the object to vibrate. The patterns of vibration are different depending on the characteristic of the objects. However, they include enough information to know the characteristics.

Abe Davis *et al.* proposed a method that classifies the characteristics of vibrating objects in videos [1]. They also showed that it is possible to estimate the movement of the force-added object by exploring vibration patterns [2]. Justin G. Chen *et al.* showed that it is possible to conduct nondestructive inspection of buildings using video of buildings shaken by the wind [3].

Some research has been carried out on the recovering of sounds by extracting the vibration of objects. This research is expected to be applied in the field of surveillance and security because sounds can be recorded from relatively far away. Before, laser microphones were proposed to extract the vibration of objects from a distance. The basic laser microphone records the phase of a reflected laser. A laser Doppler vibrometer measures the Doppler shift of the reflected laser to determine the velocity of the reflecting surface [4]. Both types of laser microphones can recover high quality sound from a distance. However, it depends on the precise positioning of a laser and receiver, as well as having a surface with the appropriate reflectance.

Abe Davis *et al.* proposed a method to recover a sound from a high-speed video [5]. This method has some advantages that it does not depend on active illumination, and it does not rely on additional sensors or detection modules other than a high-speed video camera. They also show how sound may be recovered from regular consumer cameras with standard frame-rates. In our previous work [6], we recovered the sound from a video of an object's subtle motion in the presence of large motions by using momentary phase variations.

In this paper, we propose a method that recovers sound by considering the vibration direction of the object. In the conventional method, sound is recovered only from vibrations in a predetermined direction. So it is considered that the better sound can be obtained by recovering it based on the direction of the largest vibration. First, the sound is recovered in a certain direction (initial direction) and then the sound is recovered in the direction that has the largest correlation with the first recovered sound. By repeating this process, the sound is recovered in the direction of the largest vibration. We recovered sounds from several objects in videos and evaluated the effectiveness of our method.

II. Sound Recovery

A. Extraction of Small Displacement

For simplicity, we consider the case of a 1D image intensity profile $f(x)$. By Fourier series decomposition, $f(x)$ is represented as a sum of complex sinusoids. The displacement of the image affects the phase only. Thus the image profile displaced by the displacement function $\delta(t)$ is represented as

$$f(x+\delta(t)) = \sum_{\omega=-\infty}^{\infty} A_\omega e^{i\omega(x+\delta(t))}. \quad (1)$$

ω represents the spatial angular frequency and A_ω represents the amplitude of each component.

Therefore, we can obtain the displacements of images using the phase difference.

B. Conventional Sound-Recovery Method

Neal Wadhwa *et al.* extract local small changes from phase variations in the complex steerable pyramid [7], and uses them

978-1-5386-6857-3/18/$31.00 ©2018 IEEE
DOI 10.1109/ISM.2018.00-17

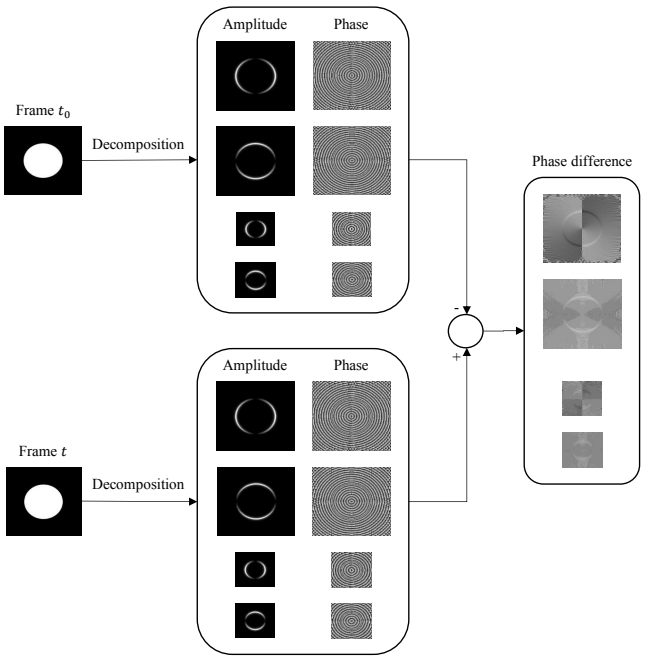

Fig. 1. Procedure of extraction of phase variation

to magnify the local subtle motions [8]. A complex steerable pyramid is a filter bank that decomposes an image into complex-valued spatial sub-bands corresponding to a different scale r and an orientation θ. This process is performed in the frequency domain on the input image. First, the input image is applied a high-pass filter, and then the rest is applied a low-pass filter. The middle band is applied an oriented filter. The sub-bands are inversely transformed and output as complex images. The removed low-frequency bands are subsampled, and this process is recursively repeated.

The complex image $I_{r,\theta}$, which represents the sub-band of scale $r = 1, \ldots, n$ and orientation $\theta = 1, \ldots, m$, is decomposed into amplitude $A_{r,\theta}(\mathbf{x})$ and phase $\phi_{r,\theta}(\mathbf{x})$ by using Euler's formula as

$$A_{r,\theta}(\mathbf{x}) = \sqrt{Re\left(I_{r,\theta}(\mathbf{x})\right)^2 + Im\left(I_{r,\theta}(\mathbf{x})\right)^2}, \tag{2}$$

$$\phi_{r,\theta}(\mathbf{x}) = \arctan \frac{Im\left(I_{r,\theta}(\mathbf{x})\right)}{Re\left(I_{r,\theta}(\mathbf{x})\right)}. \tag{3}$$

$\mathbf{x}$ represents the position in the image.

The phase difference $\phi^v_{r,\theta}(\mathbf{x}, t)$ between a frame t and a reference frame t_0 is calculated for all t to obtain the phase variation as

$$\phi^v_{r,\theta}(\mathbf{x}, t) = \phi_{r,\theta}(\mathbf{x}, t) - \phi_{r,\theta}(\mathbf{x}, t_0). \tag{4}$$

Fig. 1 shows an example of phase variation extraction. In Fig. 1, the one-pixel displacement of the disc image is extracted. The phase difference image represents the displacements of each pixel. In textureless regions, noise factors for phase tend to increase. Therefore, the single motion signal $\Phi_{r,\theta}(t)$ of the sub-band at frame t is calculated as the spatial average of phase variations weighed by its squared amplitude as

$$\Phi_{r,\theta}(t) = \sum_{\mathbf{x}} A_{r,\theta}(\mathbf{x}, t)^2 \phi^v_{r,\theta}(\mathbf{x}, t), \tag{5}$$

because the amplitude gives the strength of texture.

Finally, single motion signals are aligned temporally to relate to each other, and they are combined into the recovered signal $\hat{s}(t)$ to strengthen each signal as

$$t_i = \arg\max_{t_i} \Phi(r_0, \theta_0, t)^T \Phi(r_i, \theta_i, t - t_i), \tag{6}$$

$$\hat{s}(t) = \sum_i \Phi(r_i, \theta_i, t - t_i). \tag{7}$$

Moreover, the recovered sound is further processed for the denoising. To remove the noise at the lower frequencies, a high-pass Butterworth filter is applied to the recovered signal. To improve the signal more, some denoising methods [9], [10] are applied to it.

In this method, the orientation θ is predetermined. If $\theta = 2$, vertical and horizontal vibration is obtained. However, the vibration direction of the object varies depending on the conditions. Therefore, it is considered that more accurate sound recovery can be achieved by processing the sound in the correct vibration direction.

III. Proposed Method

In this work, in order to estimate the vibration direction for the whole object from phase-difference images, the Fourier transform is applied to raw images instead of the complex steerable pyramid that estimates the vibration for each pixel.

First, the Fourier transform is applied to raw images, where the obtained amplitude $A(\omega_x, \omega_y, t)$ and phase $\phi(\omega_x, \omega_y, t)$ represent the strength and position (movement) of each frequency component, respectively. Therefore, the movement of each direction can be obtained as the product of the phase difference between the previous frame and the current frame and the square of the time average of the amplitude.

$$\phi_v(\omega_x, \omega_y, t) = \phi(\omega_x, \omega_y, t) - \phi(\omega_x, \omega_y, t - 1) \tag{8}$$

$$s(\omega_x, \omega_y, t) = \bar{A}(\omega_x, \omega_y)^2 \phi_v(\omega_x, \omega_y, t) \tag{9}$$

It is converted into the movement from the first frame by adding the movement up to the current time.

$$s'(\omega_x, \omega_y, t') = \sum_{t=t_0}^{t'} s(\omega_x, \omega_y, t) \tag{10}$$

In order to obtain the movement of the whole image in a specific direction, the signals of each frequency component are weighted by the oriented filter $g(\theta, \omega_x, \omega_y)$ and added together. Here, θ represents the angle of vibration and has a value in the range from π to $-\pi$. The sound is recovered as follows (Fig. 2):

$$g(\theta, \omega_x, \omega_y) = \cos(\arctan(\omega_y/\omega_x) - \theta) \tag{11}$$

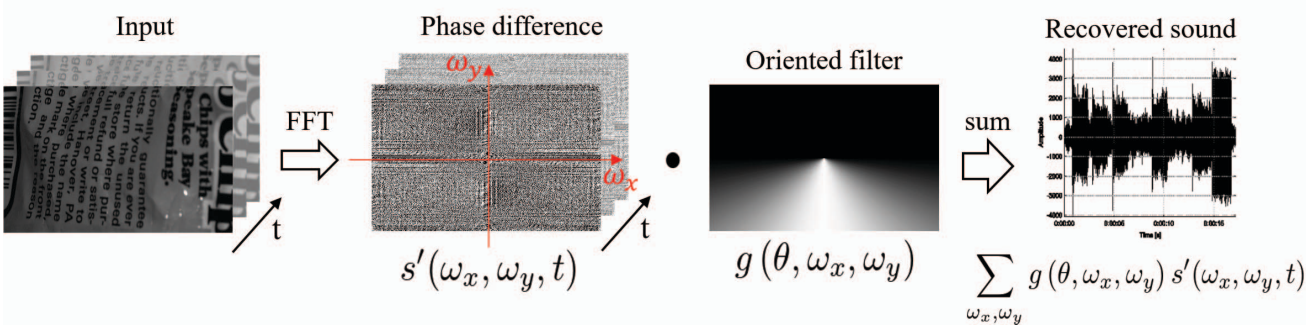

Fig. 2. Procedure of recovering the sound for an direction

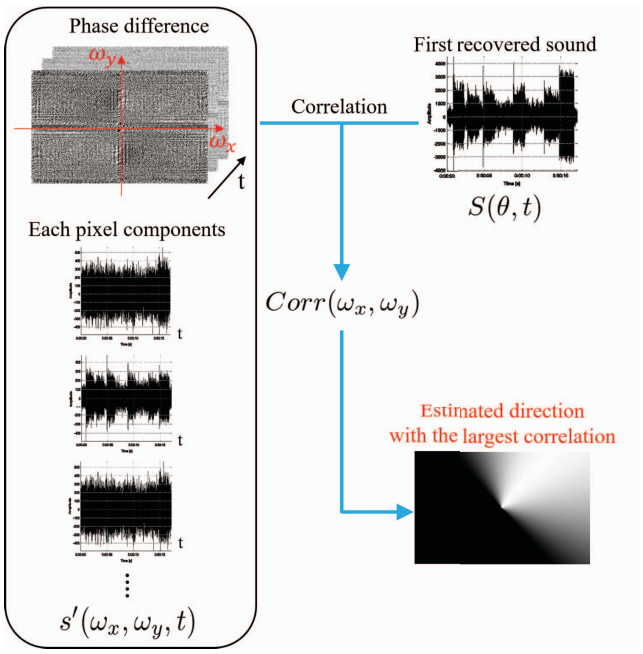

Fig. 3. Estimation of vibration direction

$$S(\theta, t) = \sum_{\omega_x,\omega_y} g(\theta, \omega_x, \omega_y)\, s'(\omega_x, \omega_y, t) \qquad (12)$$

A high-pass Butterworth filter is applied to the recovered sound like the conventional method.

In order to obtain the direction with the largest vibration, first, the sound is recovered for two directions of $\theta = 0, -\pi/2$, which represent horizontal and vertical vibration. Next, we calculate the correlation between the signal for each frequency component and the two recovered sounds. Then, the angle corresponding to the largest correlation is obtained as the vibration direction. The procedure is shown in Fig. 3. By repeating this until the difference between the two angles becomes 0 or π, we obtain the angle $\arctan(\hat{\omega}_y/\hat{\omega}_x)$, which can recover the best sound. Correlation between $S(\theta, t)$ and $s'(\omega_x, \omega_y, t)$ was calculated in the length of l and then $(\hat{\omega_x}, \hat{\omega_y})$

Fig. 4. Sample frames of videos

is chosen to maximize it.

$$(\hat{\omega}_x, \hat{\omega}_y) = \arg\max_{\omega_x,\omega_y} \sum_{t=0}^{l-1} S(\theta, t) s'(\omega_x, \omega_y, t) \qquad (13)$$

IV. Experiments

A. Experimental setup

Plastic bags of different sizes (shown in Fig. 4) are filmed in the same environment. The objects are illuminated with additional photography lamps and filmed from about 10 cm away using a high-speed camera. Sound is played by the loudspeaker at volumes over 100 dB. The loudspeaker is placed over 30 cm away from the object, and its direction is at a right angle to the direction of the camera. The chirp signal is used as the sound source. The video frame rate is 2,200 Hz, and the resolution is 256×256 pixels. Recovered sounds are denoised using a high-pass Butterworth filter with a cut-off of 80 Hz.

B. Experimental results

Fig. 5 shows the spectrograms of the recovered sounds from object 1. Fig. 6 shows the correlation between the signals for each frequency component and the recovered sound of object 1. The upper row "before" shows the first recovered sound for the initial direction, and the lower row "after" shows the recovered sound after estimating the vibration direction. $\theta = 0$ means horizontal vibration and $\theta = -\pi/2$ means vertical vibration.

Table I shows the segmental SNR (SSNR) [11] of the sounds recovered by our method and the conventional method. SSNR is improved, especially for the sound recovered from the horizontal vibration ($\theta = 0$). As shown in the left half of

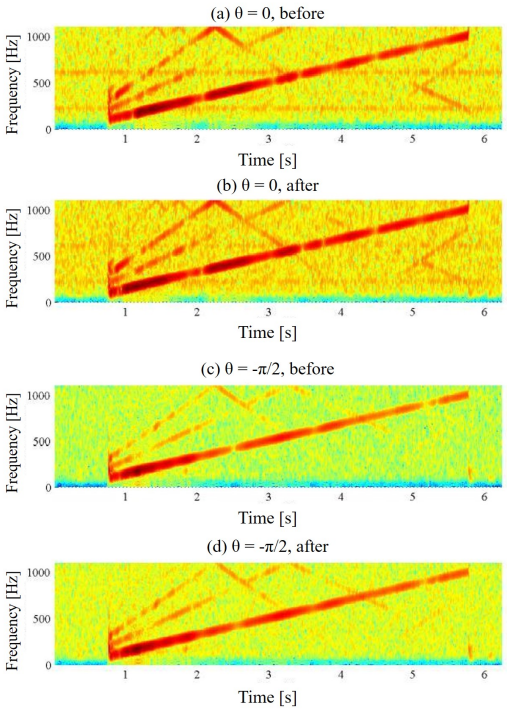

Fig. 5. Sound recovered from object 1

TABLE I
SSNR OF RECOVERD SOUND FOR THE CHIRP SIGNAL

SSNR [dB]	Object 1	Object 2	Object 3
Conventional method	0.6114	1.0259	0.8882
Proposed method ($\theta = 0$)	0.8388	1.0309	0.589
Proposed method ($\theta = -\pi/2$)	1.5857	2.0808	1.15
Proposed method ($\theta = 0$, estimated)	1.8304	1.7705	1.632
Proposed method ($\theta = -\pi/2$, estimated)	1.7007	1.6234	1.1617

Fig. 6, after estimating the direction of the largest vibration, the correlations for the horizontal signal and the recovered sound have changed. It is considered that the vibration direction is perpendicular to the boundary between the white region and the black region. That is, the sound first recovered from the horizontal vibration is improved by estimating the vibration direction.

For object 2, the SSNR of the sound recovered from the vertical vibration ($\theta = -\pi/2$) decreased. It is possible that the amplitude weight used for recovering the sound affects the result because the amplitude weight depends on the direction of the pattern of the object.

V. Conclusions

We proposed the sound-recovery method considering the vibration direction of the object. It has been shown that it is effective in recovering the sound from the direction with the largest vibration. In future research, we will investigate the efficiency of our approach on various objects.

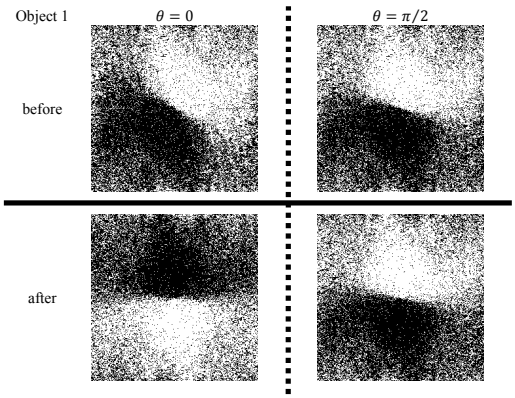

Fig. 6. Correlation between the frequency components and recovered signals (object 1)

Acknowledgment

This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2) and JSPS KAKENHI (Grant No. JP17H01995).

References

[1] Abe Davis *et al.*, "Visual Vibrometry: Estimating Material Properties from Small Motion in Video," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR),2015.
[2] Abe Davis *et al.*, "Image-space modal bases for plausible manipulation of objects in video," ACM Transactions on Graphics,34(6),239:1-239:7,2015.
[3] Justin G. Chen *et al.*, "Video camera-based vibration measurement for Condition Assessment of Civil Infrastructure," NDT-CE International Symposium Non-Destructive Testing in Civil Engineering,15-17,2015.
[4] Rothberg, S., Baker, J., and Halliwell, N. A., "Laser vibrometry: pseudo-vibrations," Journal of Sound and Vibration, 135 (3), 516-522, 1989.
[5] Abe Davis *et al.*, "The Visual Microphone: Passive Recovery of Sound from Video," ACM Transactions on Graphics, 33 (4), 79:1-79:10, 2014.
[6] Yusuke Yasumi *et al.*, "Visual Sound Recovery Using Momentary Phase Variations," The 23rd International Workshop on Frontiers of Computer Vision, 2-6, 2017.
[7] Javier Portilla *et al.*, "A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients," International Journal of Computer Vision, 40 (1), 49-71, 2000.
[8] Neal Wadhwa *et al.*, "Phase-Based Video Motion Processing," ACM Transactions on Graphics, 32(4), 80:180:10, 2013.
[9] Steven F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Transactions on Acoustics, Speech and Signal Processing, 27(2), 113-120, 1979.
[10] Lu, Yang and Loizou, P. C., "A geometric approach to spectral subtraction," Speech Communication, 50(6) , 453-466, 2008.
[11] John H. L. Hansen and Bryan L. Pellom, "An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms," Proceedings of the International Conference on Speech and Language Processing, 2819-2822, 1998.