Visual Sound Recovery Using Momentary Phase Variations

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

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

Yasuo Ariki
Graduate School of System Informatics,
Kobe University,
Kobe, Japan
Email:ariki@kobe-u.ac.jp

***Abstract*—Because sound is basically a pressure fluctuation of the air, sounds cause minute vibrations that are usually invisible on an object. In the field of computer vision, we can deal with the subtle motions that our eyes cannot capture. Therefore, using a computer vision approach, we can recover sound from the subtle motion on the object in a video [1]. However, when large motions of the object occur in a video due to blowing wind or camera shaking, it may cause several problems for visual sound recovery. In this paper, we attempt to recover the sound from an object's subtle motion in the presence of large motions. A new method using momentary phase variations is proposed to resolve this problem. Its effectiveness is confirmed by experiments on recovering sound from a plastic shopping bag blowing in the wind.**

I. Introduction

In our lives, we can notice various motions of objects in our field of sight; for example, a running child, a shaking plant, a flying bird, and so on. On the other hand, there are many motions that we cannot see - conditions in which objects often move so subtly that we think they are standing still. In such cases, the object's motion includes more information than we are able to see. The invisible information, however, may be useful in recovering sound.

The maximum distance at which microphones can pick up faraway sounds clearly is about ten meters. However, the maximum distance at which cameras can see objects is much farther than microphones. Cameras also enable us to see places that we cannot approach physically. Thus, we can "hear" farther sounds than those microphones can pick up by recovering sound from images.

In this paper, a method for extracting sound from subtle motions of objects in the presence of large motions is proposed. Because of wind, camera shaking or motion, an object's motion is often visible in a real environment. In the case of large motions, there are also subtle motions taht can be used to recover sounds. Our method employs a momentary phase difference in the object, and recovers sound by integrating the difference. Fig. 1 shows the process of our method, which improves on the research work [1] in the process of extracting signals and calculating single motion signals. This is described in detail in Section 3.

II. Related research works

Recently, some studies have dealt with subtle invisible motions in video images, video magnification to study subtle motions and make these motions visible.

Hao-Yu Wu et al. [2] makes subtle motions or invisible color changings visible by amplifying pixel-wise temporal brightness changes.

Neal Wadha et al. [3] extracts local subtle changes from phase variations in the complex steerable pyramid proposed by Simoncelli et al. [4], and uses these to magnify the local subtle motions.

Mohamed A. Elgharib et al. [5] combines the tracking of a region of interest using optical flow or iterative stabilization with phase-based video magnification, to magnify subtle motions in the presence of large motions. This work also extracts subtle motions in large motions, but it requires more calculations than our method. In addition, it does not aim to recover sound.

Next, we will show the conventional method, and how it extracts subtle motions from a video.

A. Displacement of an image

By using Fourier series decomposition, the image profile, $f(\mathbf{x})$, is represented as a sum of complex sinusoids

$$f(\mathbf{x}) = \sum_{\omega=-\infty}^{\infty} A_w e^{iw\mathbf{x}}, \tag{1}$$

where $\mathbf{x}$ is the vector for each pixel.

This means that the displacement of the image affects the phase only. Therefore, the image profile displaced by the function $\delta(t)$ is represented by

$$f(\mathbf{x} + \delta(t)) = \sum_{\omega=-\infty}^{\infty} A_w e^{iw(\mathbf{x}+\delta(t))}. \tag{2}$$

Thus, we can obtain the displacements of images using the difference in phases.

B. Conventional method

In [1], the complex steerable pyramid [4] is used to obtain the phase variation. This is a filter bank, and it decomposes the image into complex-valued spatial subbands corresponding to

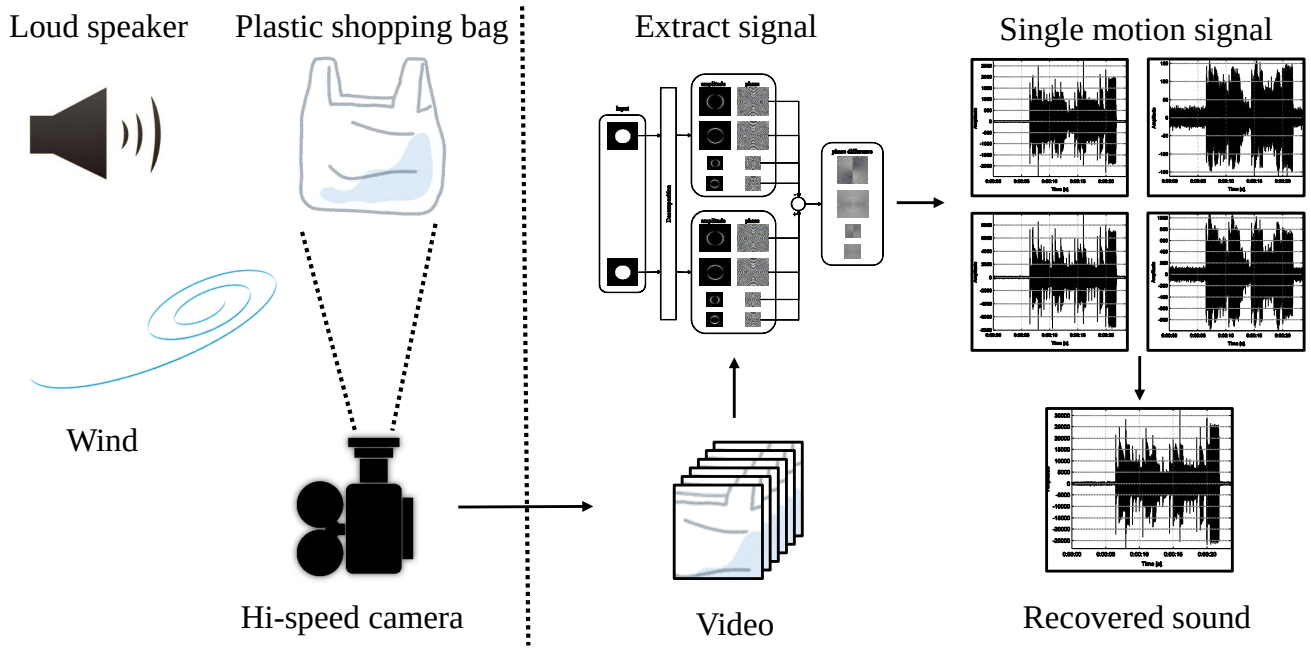

Fig. 1: Process of sound recovery

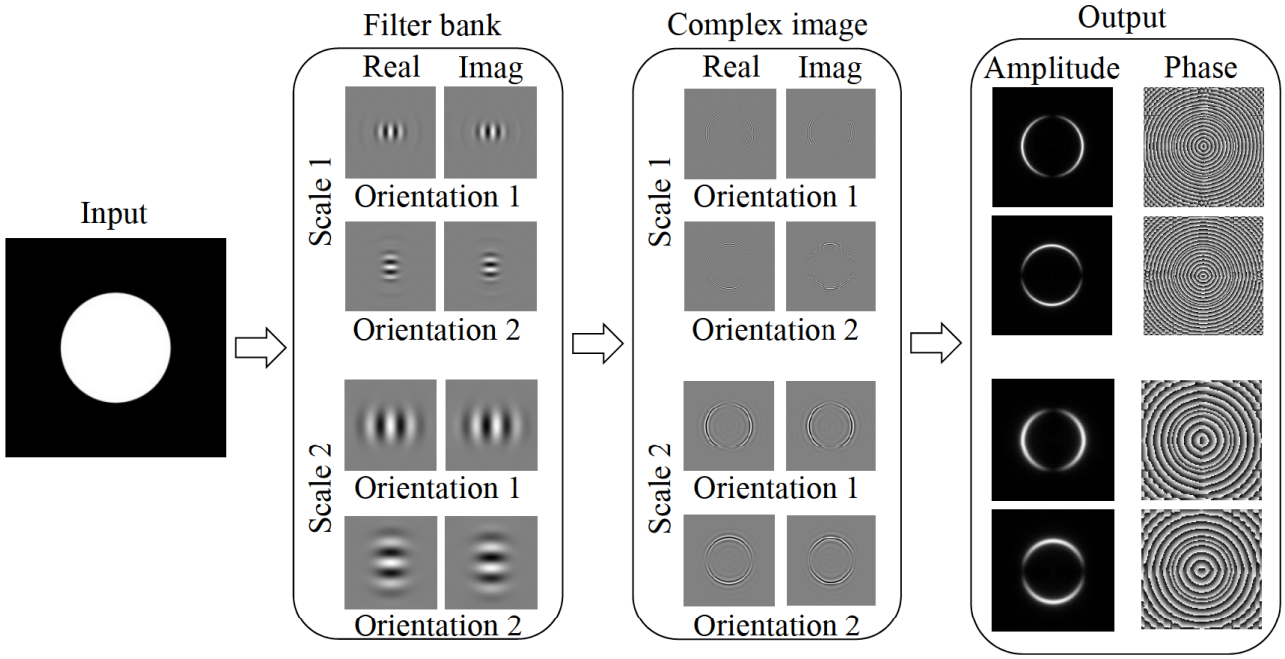

Fig. 2: Decomposition of disc image

a different scale r and an orientation θ. Thus, by using Euler's formula, the subband is expressed as

$$A(r,\theta,\mathbf{x})e^{i\phi(r,\theta,\mathbf{x})}. \tag{3}$$

$A(r,\theta,\mathbf{x})$ is the amplitude of the subband corresponding to r and θ, and $\phi(r,\theta,\mathbf{x})$ is the phase of the subband. Fig. 2 shows the decomposition process of a disc image. The image is decomposed into four subbands corresponding to two scales and two orientations.

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

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

In textureless regions, noise factors for phase tend to increase. Therefore, because the amplitude gives the strength of texture, the single motion signal $\Phi(r,\theta,t)$ of the subband at frame t is calculated as the spatial average of phase differences weighed by its squared amplitude as follows:

$$\Phi(r,\theta,t) = \sum_{\mathbf{x}} A(r,\theta,\mathbf{x},t)^2 \phi_v(r,\theta,\mathbf{x},t). \tag{5}$$

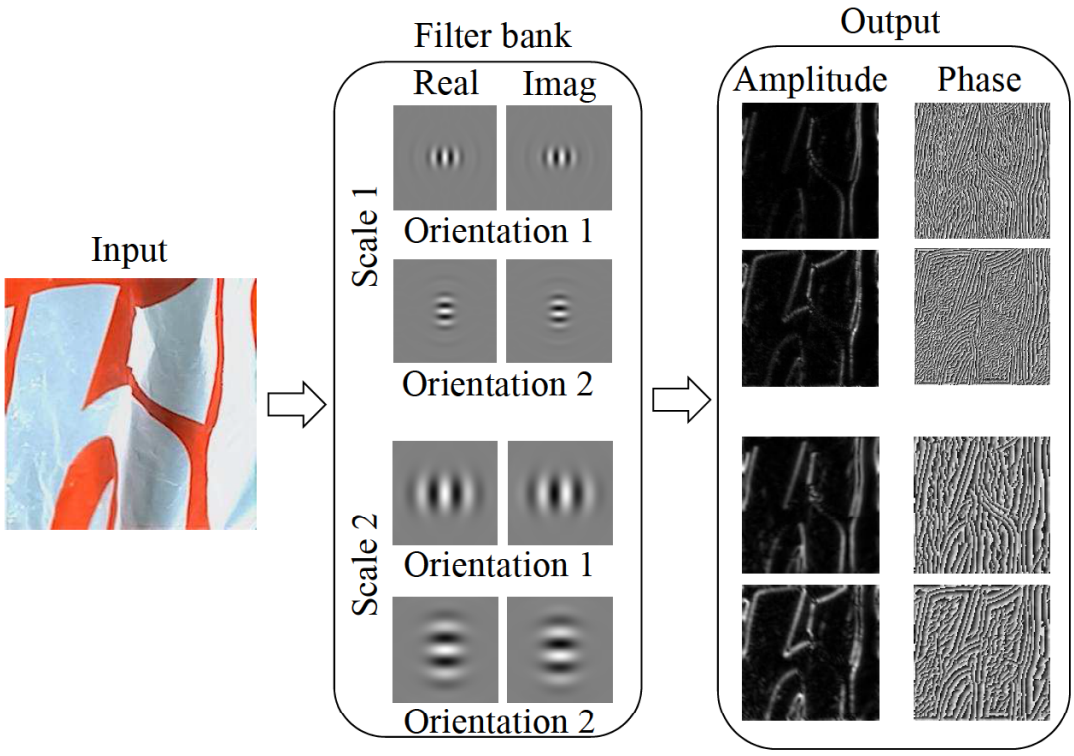

Fig. 3: Decomposition of a frame of input videos

Finally, single motion signals are aligned temporally to relate to each other, and they are combined into the recovered signal to strengthen each signal.

Moreover, for the denoising, the recovered signal is further processed. To remove high-energy noise in the lower frequencies, a high pass Butterworth filter is applied to the recovered signal. To improve the signal further, a denoising method [6], [7] is applied to it.

III. PROPOSED METHOD

In the conventional method, it supposed that object motions are affected only by sound, and they are very subtle. Therefore, when object motions become larger because of wind or some other cause, it is difficult to extract motions well. In this paper, we suppose that the momentary difference is small because of the use of a high-speed camera, and a simple method is introduced to resolve it.

First, the momentary difference is derived from the phase difference between frame t and just before frame $t-1$ as

$$\phi_v\left(r,\theta,\mathbf{x},t\right)=\phi_{r,\theta}\left(\mathbf{x},t\right)-\phi_{r,\theta}\left(\mathbf{x},t-1\right). \tag{6}$$

Then, the signal motion signal $\Phi\left(r,\theta,t'\right)$ is updated from the second frame to the last frame using the equation

$$\Phi\left(r,\theta,t\right)=\Phi\left(r,\theta,t\right)+\Phi\left(r,\theta,t-1\right). \tag{7}$$

However, because it is supposed that wind and other causes of large motions have a lower frequency than subtle motions caused by sound, it is expected that most of them are removed by applying a high pass Butterworth filter to each single motion signal before updating.

IV. EXPERIMENTS

A. Experimental setup

A plastic shopping bag is used as an object, and it is illuminated using additional photography lamps. Sound is played by loudspeaker at volumes over 100 dB. The loudspeaker is over 30 cm away from the object, and its direction is at a right angle to the camera's direction.

The video frame rate is 2,200 Hz, with resolutions of 256×256 pixels. Denoising methods are a high-pass Butterworth filter with a cut-off of 55 Hz and spectral subtraction [6]. A hi-speed camera is over 10 cm far from the object.

The complex steerable pyramid [4] is used to obtain the phase variation, where the scale is four and the orientation is two. Videos are taken in windy and windless environments. The wind is produced using a Japanese folding fan. It is weak enough not to distort the object, but the motion of the object is clearly visible in a video.

Fig. 3 shows the decomposition of a frame in the input video. In this figure, only subbands corresponding to two scales and two orientations are shown, but all 8 subbands are used for the process of the sound recovery.

B. Experimental results

Fig. 4 shows the spectrogram of the recovered sound from a video taken in the windy environment, and Fig. 5 shows the

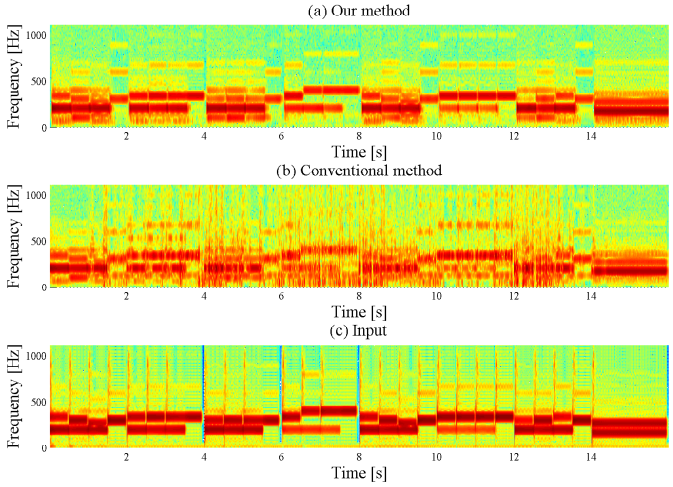

Fig. 4: “Mary had a little lamb” in a windy environment

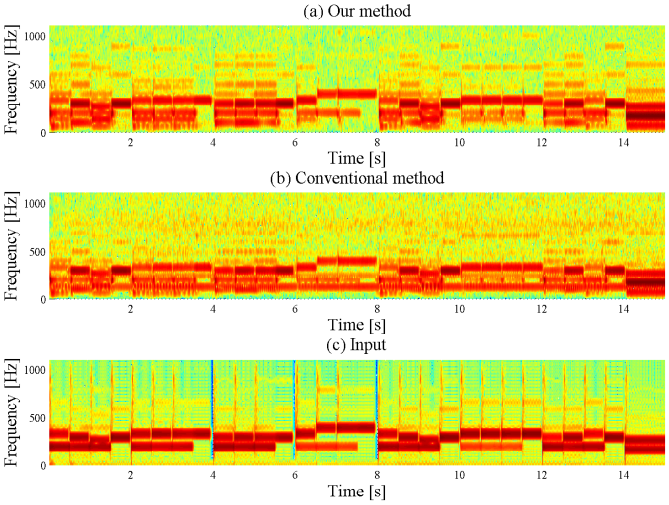

Fig. 5: “Mary had a little lamb” in a windless environment

non-windy results. These show that the quality of recovered sound for both methods is about the same if the object is not experiencing large motions caused by wind. But when there is wind, from 2 to 13 seconds, the noise in the sound recovered by the conventional method increases. The other signals are almost the same as the recovered sound without wind. As a result, when there is wind, the proposed method reduces the wind-induced noise.

Objective tests were carried out using the Spectrum Distortion (SD).

$$SD = \sqrt{\frac{1}{I}\sum_{i=1}^{I}\left(20\log\frac{|S(f_i)|}{|\hat{S}(f_i)|}\right)^2} \tag{8}$$

where S and $\hat{S}$ denotes the original spectrum and the recovered sound spectrum, respectively.

Table I shows the SD in a windy and windless environments. As shown in this table, the results of the experiment indicate that the proposed method provides better performance of the SD in comparison with the conventional method.

TABLE I: Spectral Distortion [dB]

	conventional	proposed
windy	14.46	10.92
windless	16.69	14.01

V. Conclusion

We proposed a simple method that can reduce the noise caused by large motions, and from our experimental results, it has been shown that the use of the momentary phase variations provides better sound recovery performance in wind. In future research, we will continue to investigate how to find the appropriate length between the frame and the reference frame.

Acknowledgment

This work was supported in part by PRESTO, JST.

References

[1] Abe Davis *et al.*, “The Visual Microphone: Passive Recovery of sound from Video,” ACM Transactions on Graphics, 33(4), pp. 79:1-79:10, 2014.

[2] Hao-Yu Wu *et al.*, “Eulerian Video Magnification for Revealing Subtle Changes in the World,” ACM Transactions on Graphics, 31(4), pp. 65:1-65:8, 2012

[3] Neal Wadhwa *et al.*, “Phase-Based Video Motion Processing,” ACM Transactions on Graphics, 32(4), 2013.

[4] Javier Portilla *et al.*, “A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients,” International Journal of Computer Vision 40(1), pp. 49-71, 2000.

[5] Mohamed A. Elgharib *et al.*,“Video Magnification in Presence of Large Motion,” Computer Vision and Pattern recognition CVPR, pp. 4119-4127, 2015.

[6] Steven F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 27, no. 2, pp. 113-120, 1979.

[7] Philipos C. Loizou, “Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum,” Speech and Audio Processing, IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 857-869, 2005.