Sound Recovery Using Vibration Modes of the Object in a Video

Yohei Fuse, Yusuke Yasumi and Tetsuya Takiguchi
Graduate School of System Informatics, Kobe University, Japan
E-mail: y.fuse@stu.kobe-u.ac.jp, yasumi@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp

***Abstract*—When a sound hits an object, it causes the surface of the object to vibrate. Some research has been carried out on the recovering of sounds by extracting the vibrations seen on video images. This research is expected to be applied in the field of surveillance and security because sounds can be recorded from relatively far away. The vibration of objects due to sound is so fast and minute that it is invisible. However, it is possible to observe such changes in objects by using the high-speed video as the movement of each pixel by using a complex steerable pyramid. In the conventional method, the movements of all pixels are added together to recover the sound. So it is possible that some noise source vibrations are mixed because there are some pixels that move independently of the sound source being focused upon. In this paper, we propose a sound recovery method focusing on the vibration modes of the object associated with the frequency. The vibrating parts of objects are different depending on the material, shape and frequency. The vibration is composed of some normal vibrations, and each has different loops and nodes. We confirm which part of the object is vibrating for each frequency of the sound, and recover the sound using a filter based on the response of the object. Which part is vibrating is confirmed from the amplitude response of each pixel when the signal of that frequency is the largest. This response and the reliability of the signal of each pixel are multiplied to each pixel as a filter. We recovered sounds from several objects in videos and ascertained the effectiveness of the method.**

I. Introduction

When a sound hits an object, it causes the surface of the object to vibrate. While the patterns of movement are different depending on the characteristic of the objects, they include enough information to understand 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 non-destructive 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 due to sound. This research is expected to be applied in the field of surveillance and security because sounds can be recorded from relatively far away. To extract the vibration of objects from a distance, laser microphones were proposed. 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 long distance. However, it depends on the precise positioning of a laser and receiver, as well as having a surface with the appropriate reflectance. Zalevsky *et al.* address these limitations by using an out-of-focus, high-speed camera to record changes in the speckle pattern of reflected laser light [5].

Abe *et al.* proposed a method that recovers sound from a high-speed video [6]. This technique does not depend on active illumination, and 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 their method, the movements of each pixel in a video frame are added together. As a result, some noise in the video that is not related to the sound source that is being recorded may be added to the mix. In our previous work [7], we recovered the sound from an object ' s subtle motion in the presence of large motions using momentary phase variations.

In this paper, we propose a method that recovers sound by considering the vibration modes of the object that are associated with the frequency. The object's vibration is composed of some normal vibrations (vibration modes), and each vibration mode has loops and nodes. The vibrating parts of objects differ depending on the material, shape and frequency of the vibrating parts. We confirm which part of the object is vibrating for each frequency of the sound, and recover the sound using a filter based on the response of the object. Which part is vibrating is confirmed from the amplitude response of each pixel when the signal of that frequency is the largest. This response and reliability of the signal of each pixel are multiplied to each pixel as a filter. We recovered sounds from several objects in videos and evaluated the effectiveness of our method.

II. Sound recovery

A. Displacement extraction

For simplicity, we consider the case of a 1D image intensity profile $f(x)$. Using Fourier series decomposition, $f(x)$ is represented as a sum of complex sinusoids

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

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

$$f(x+\delta(t)) = \sum_{\omega=-\infty}^{\infty} A_\omega e^{i\omega(x+\delta(t))}. \quad (2)$$

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

B. Complex steerable pyramid

In this paper, we use a complex steerable pyramid [8], which is a technique that extracts a small change in the image.

Neal Wadhwa *et al.* extract local small changes from phase variations in the complex steerable pyramid, and uses them to magnify the local subtle motions [9]. They also enable real-time processing by using a Riesz Pyramid, which performs as well as a complex steerable pyramid [10]. Mohamed A. Elgharib *et al.* combine 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 [11]. 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. Fig. 1 shows the procedure of image decomposition using the filter bank. This process is performed in the frequency domain on the input image. First, the input image is run through a high-pass filter, and then the rest is run through a low-pass filter. The middle band is run through 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. Fig. 2 shows an example of the image applied this filter bank.

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

$$A_{r,\theta}(\mathbf{x}) = \sqrt{Re\left(I_{r,\theta}(\mathbf{x})\right)^2 + Im\left(I_{r,\theta}(\mathbf{x})\right)^2}, \quad (3)$$

$$\phi_{r,\theta}(\mathbf{x}) = \arctan \frac{Im\left(I_{r,\theta}(\mathbf{x})\right)}{Re\left(I_{r,\theta}(\mathbf{x})\right)}. \quad (4)$$

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

C. Conventional sound recovery method

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

$$\phi^v_{r,\theta}(\mathbf{x},t) = \phi_{r,\theta}(\mathbf{x},t) - \phi_{r,\theta}(\mathbf{x},t_0). \quad (5)$$

Fig. 3 shows an example of phase variation extraction. In Fig. 3, 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

$$\Phi_{r,\theta}(t) = \sum_{\mathbf{x}} A_{r,\theta}(\mathbf{x},t)^2 \phi^v_{r,\theta}(\mathbf{x},t), \quad (6)$$

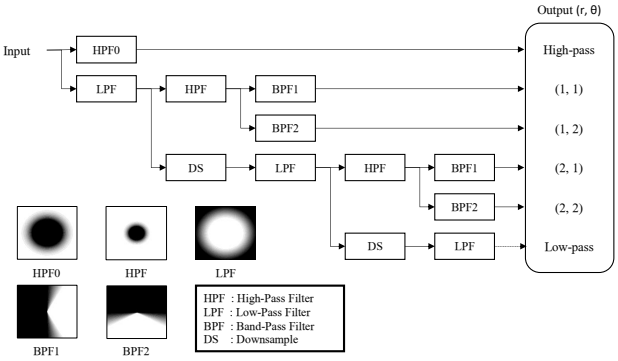

Fig. 1. Procedure of image decomposition using complex steerable pyramid $(r,\theta) = (2,2)$

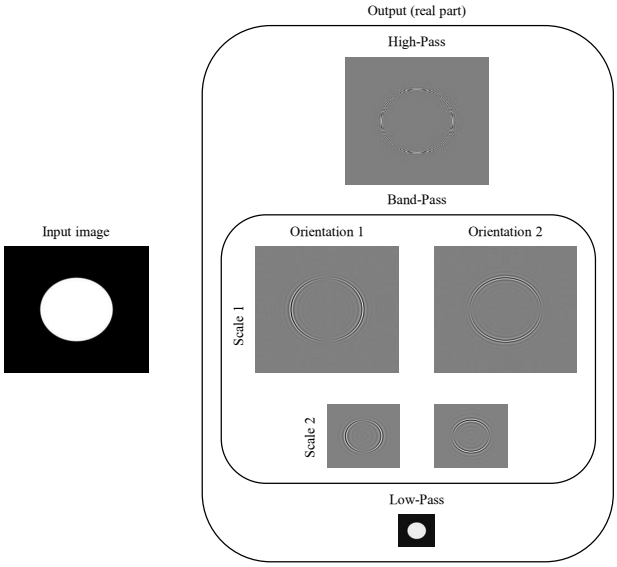

Fig. 2. Example of image decomposition

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

$$t_i = \arg\max_{t_i} \Phi(r_0,\theta_0,t)^T \Phi(r_i,\theta_i,t-t_i), \quad (7)$$

$$\hat{s}(t) = \sum_i \Phi(r_i,\theta_i,t-t_i). \quad (8)$$

Moreover, the recovered signal is further processed for the denoising. To remove high-energy noise in the lower frequencies, a high-pass Butterworth filter is applied to the recovered signal. To improve the signal even more, a denoising method [12], [13] is applied to it.

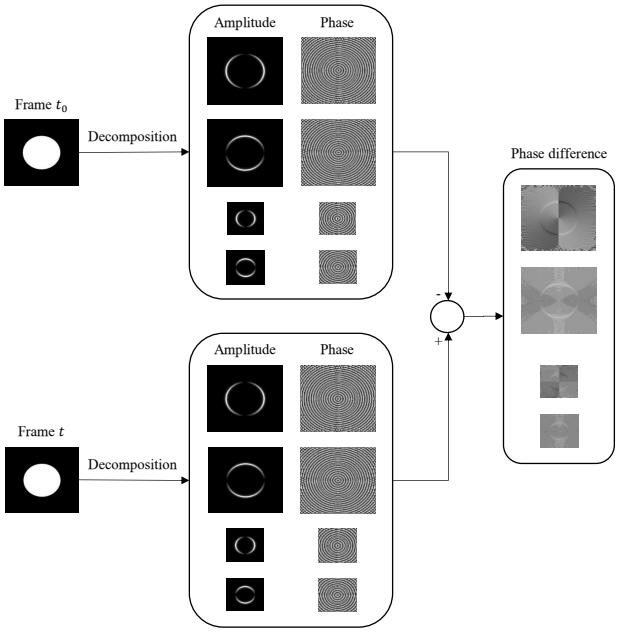

Fig. 3. Procedure of extraction of phase variation

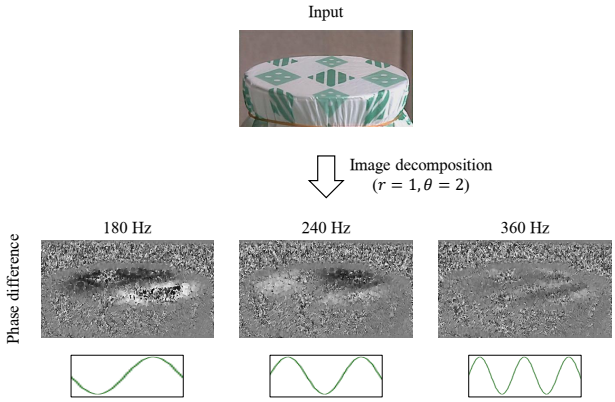

Fig. 4. Phase difference at each frequency

III. Proposed Method

A. Vibration modes

An object has parts that are easy or hard to vibrate. The parts depend on the frequency because the vibration modes that compose the object's vibration have different loops and nodes.

Fig. 4 shows an example in which a cup covered with nylon is hit by sounds of several frequencies. The images represent phase differences, where the white pixel means that the phase difference is π and represents upward displacement. The black pixel means $-\pi$ and represents downward displacement. We can see that different parts vibrate differently depending on the frequency.

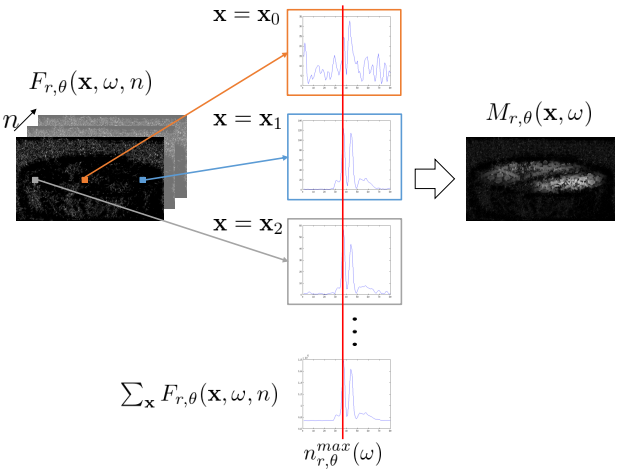

Fig. 5. Procedure for obtaining the vibration modes

B. Filtering based on vibration modes

We extract the displacement of each pixel in the input frame t by using a complex steerable pyramid as well as the conventional method. Then, we define that the movement of the object $s_{r,\theta}(\mathbf{x}, t)$ is the phase difference $\phi^v_{r,\theta}(\mathbf{x}, t)$ of frame t and position $\mathbf{x}$.

First, we determine the frame $n^{max}_{r,\theta}(\omega)$ where the signal of each frequency is the largest. We perform Short Time Fourier Transform (STFT) on the movements of each pixel in frame n, and obtain the spectrum of phase variation $F_{r,\theta}(\mathbf{x}, \omega, n)$ as

$$F_{r,\theta}(\mathbf{x}, \omega, n) = STFT[s_{r,\theta}(\mathbf{x}, t)]. \tag{9}$$

$STFT[\cdot]$ represents Short Time Fourier Transform operation. The spectrum of phase variation and the minimum of amplitude $A^{min}_{r,\theta}(\mathbf{x})$ from a complex steerable pyramid are normalized respectively, and then multiplied. The result is considered as the power of each frequency in the spatial sub-band. The frame $n^{max}_{r,\theta}(\omega)$ gives the maximum of the power as

$$n^{max}_{r,\theta}(\omega) = \arg\max_{n} \sum_{\mathbf{x}} A^{min}_{r,\theta}(\mathbf{x}) F_{r,\theta}(\mathbf{x}, \omega, n). \tag{10}$$

Then we obtain $M_{r,\theta}(\mathbf{x}, \omega)$ which is the largest response for each frequency in each sub-band as

$$M_{r,\theta}(\mathbf{x}, \omega) = |F_{r,\theta}(\mathbf{x}, \omega, n^{max}_{r,\theta}(\omega))|. \tag{11}$$

The procedure is shown in Fig. 5.

We consider the response as a filter for the vibration modes and the amplitude minimum as a filter associated with the reliability of the phases at each pixel. These filters are normalized in the sub-bands and applied to the movement in the frequency domain. The output is inversely transformed, and the real parts of all pixels are added up in each sub-band.

Object 1 Object 2 Object 3

Fig. 6. Sample frames of videos

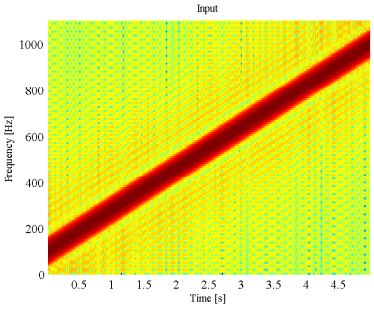

Fig. 7. Spectrogram of input signal (chirp)

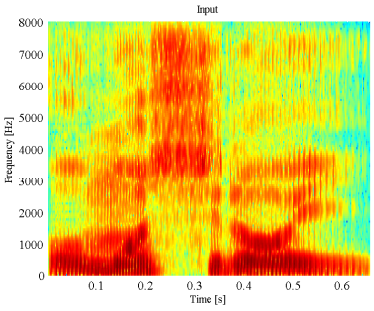

Fig. 8. Spectrogram of the utterance of /o m o sh i r o i/

IV. Experiments

A. Experimental setup

Plastic bags of different sizes (shown in Fig. 6) 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 shown in Fig. 7 is used as the sound source. The video frame rate is 2,200 Hz, with a resolution of 256×256 pixels. Denoising methods are a high-pass Butterworth filter with a cut-off of 80 Hz.

We also recover the speech shown in Fig. 8 from object 1. The video frame rate is 16,000 Hz. The recovered sound is denoised with [13] as well as with the Butterworth filter.

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	2.3520	2.1770	1.2008

TABLE II
Evaluation of recoverd sound for the utterance of /o m o sh i r o i/

	SSNR [dB]	STOI
Conventional method	-0.5806	0.5989
Proposed method	-0.4093	0.6227

We recover sounds from each sub-band, and the best of them is used as the final output.

B. Experimental result

Figs. 9-11 show the spectrograms of recovered sounds. Table I shows the Segmental SNR (SSNR) [14] of the sounds recovered by our method and the conventional method. SSNR is obtained by dividing the signal into M frames of length N and averaging the value of the SNR for each frame as

$$SSNR = \frac{10}{M}\sum_{m=0}^{M-1}\log\frac{\sum_{n=Nm}^{Nm+N-1}(y(n))^2}{\sum_{n=Nm}^{Nm+N-1}(s(n)-y(n))^2}. \quad (12)$$

$y(t)$ is the original signal and $s(t)$ is the processed signal.

Fig. 12 shows the spectrogram of the recoverd speech. Table II shows SSNR and Short Time Objective Intelligibility (STOI) [15].

In our results, there is less noise in the sound recovered by our method, and SSNR is improved from the conventional method. Sound quality is also improved in the case of speech. On the other hand, it seems that some parts of the sounds are not recovered well possibly because the vibration modes are not determined well. It may be possible to recover a better sound by improving the method of detecting an object's vibration modes.

As with the conventional method, our method also restores overtones that really do not exist. The overtones appear especially when using a sinusoidal sound source. It is considered that the overtone strength depends on the simplicity of the vibration and the object's characteristics. This needs further investigation.

V. Conclusions

We proposed the sound recovery method considering an object's vibration modes associated with the frequency. It has been shown that our method performs better sound recovery than the conventional method. In future research, we will continue to investigate how to improve the detection of an objects' vibration modes.

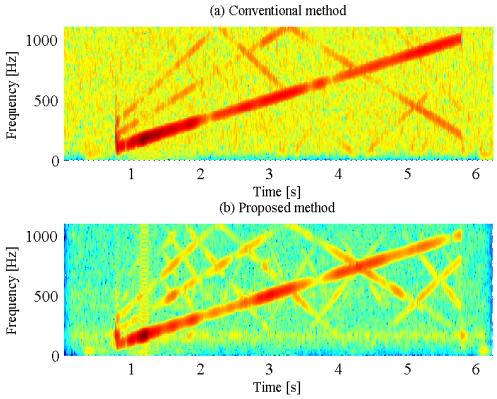

Fig. 9. Spectrogram of sound recovered from object 1

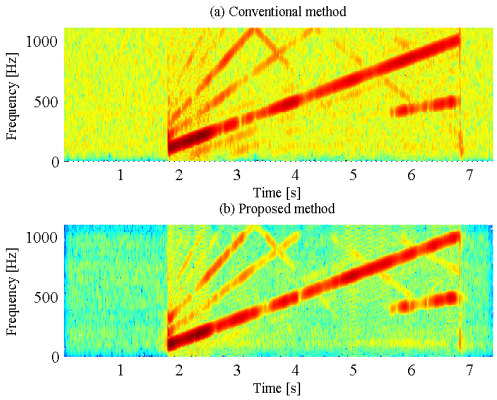

Fig. 10. Spectrogram of sound recovered from object 2

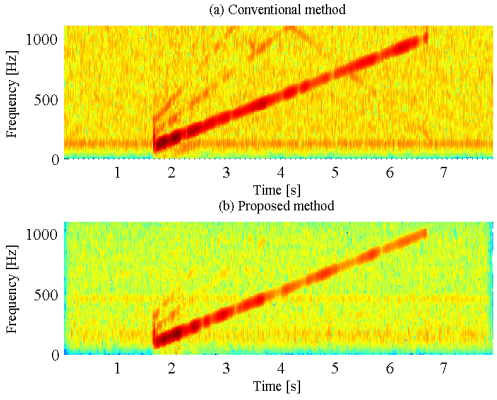

Fig. 11. Spectrogram of sound recovered from object 3

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.

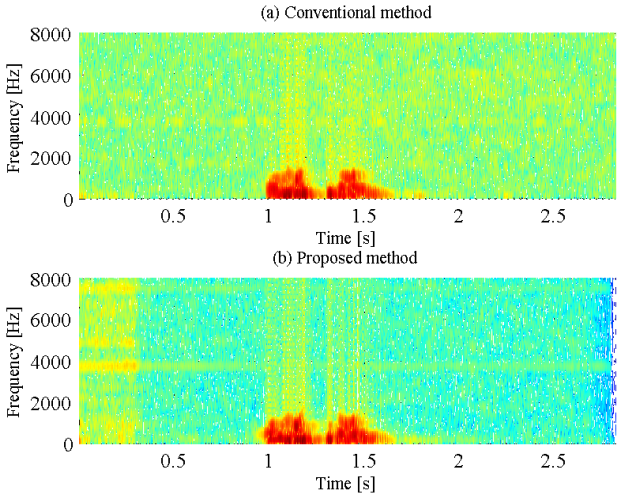

Fig. 12. Spectrogram of recovered utterance of /o m o sh i r o i/

[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] Zeev Zalevsky *et al.*, "Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern," Optics Express, 17(24), 21566-21580, 2009.
[6] Abe Davis *et al.*, "The Visual Microphone: Passive Recovery of Sound from Video," ACM Transactions on Graphics, 33 (4), 79:1-79:10, 2014.
[7] Yusuke Yasumi *et al.*, "Visual Sound Recovery Using Momentary Phase Variations," The 23rd International Workshop on Frontiers of Computer Vision, 2-6, 2017.
[8] 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.
[9] Neal Wadhwa *et al.*, "Phase-Based Video Motion Processing," ACM Transactions on Graphics, 32(4), 80:180:10, 2013.
[10] Neal Wadhwa*et al.*, "Riesz Pyramids for Fast Phase-Based Video Magnification," IEEE International Conference on Computational Photography (ICCP), 1-10, 2014.
[11] Elgharib, M.A., Hefeeda, M., Durand, F., and Freeman, W.T., "Video Magnification in Presence of Large Motions," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[12] 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.
[13] Lu, Yang and Loizou, P. C., "A geometric approach to spectral subtraction," Speech Communication, 50(6) , 453-466, 2008.
[14] 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.
[15] Taal, Cees H., Hendriks, Richard C., Heusdens, R., and Jensen, J., "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," IEEE Transactions on Audio, Speech and Language Processing, 19(7), 2125-2136, 2011.