Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Signal Processing in Acoustics
Session 1pSPb: Acoustic Feature Extraction and Characterization

1pSPb3. Extract voice information using high-speed camera

Mariko Akutsu*, Yasuhiro Oikawa and Yoshio Yamasaki

***Corresponding author's address: Department of Intermedia Art and Science, Waseda university, Shinjuku-ku, 169-8555, Tokyo, Japan, a-mariko@ruri.waseda.jp**

Conversation is one of the most important channels for human beings. To help communications, speech recognition technologies have been developed. Above all, in a conversation, not only contents of utterances but also intonations and tones include important information regarding a speaker's intention. To study the sphere of human speech, microphones are typically used to record voices. However, since microphones have to be set around a space, their existences affect a physical behavior of the sound field. To challenge this problem, we have suggested a recording method using a high-speed camera. By using a high-speed camera for recording sound vibrations, it can record two or more points within the range of the camera at the same time and can record from a distance, without interfering with the sound fields. In this study, we extract voice information using high-speed videos which capture both a face and a cervical part of the subject. This method allows recording skin vibrations which contain voices with individuality and extrapolating sound waves by using an image processing method. The result of the experiment shows that a high-speed camera is capable of recording voice information.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4800866]
Received 22 Jan 2013; published 2 Jun 2013

INTRODUCTION

Various recording methods without a microphone are proposed. With these methods, a number of microphones need not to be set for recording voices and sound. Additionally, sound can be recorded far from record point with no influences of devices. The acoustical measurement method using a laser Doppler vibrometer and the optical wave microphone are suggested for measuring sound fields [1][2][3][4]. The recording method using a high-speed camera is also proposed [5][6][7][8]. Because of its higher frame rates, high-speed cameras can record object motions including sound vibration without affected by circumstances. In addition, the high-speed camera records all area within the angle of field and its video allows to separate sounds simply and easily. To utilize these characteristics, some recording methods using high-speed camera have been studied. For example, a loudspeaker's sounds are extracted from the video of the cone paper and sound field information are extracted from the video of oil drops and dusts.

Lip-reading is also one of the way to expect words from only visual information. As tools of communication, lip-reading is mainly used by the hearing impaired. It indicates that the lips shape is important key to know the contents of conversation. On account of its noise robustness, various methods for computer lip-reading and audio-visual multi-modal speech recognition have been examined [9][10].

Moreover, speech recognition techniques have been researched by many researchers in recent years. Several services based on these technologies are provided especially for mobile terminal. Generally, computer lip-reading speech recognition converts voice sound to textual information, but to realize confortable speech communication, voice quality and intonation information are important as well as textual information. We always get a sense of speaker's intention from voice qualities and intonations.

In this paper, we try to reconstruct a speech voice include voice information like a voice quality or intonation from high-speed videos. We examine the possibility to use the method as the communication tool. To achieve recording voice include voice qualities and intonations, we use a high-speed camera. When recording voice using a high-speed camera, there are advantages of that noise robust, speaker separation and extraction voice information. At this time, we recorded speaker's lips and cervical part. From these videos, sound vibrations were extracted, and it was found that the vibrations from high-speed videos contain the same frequencies and intonations as speaker's voice. Besides, lip shape changes with words were confirmed. In the future, to reconstruct speech exactly, words and voice information will be extracted from lips shape and skin vibration, respectively.

SOUND RECORDING USING HIGH-SPEED CAMERA

High-Speed Camera

A high-speed camera is a camera, which can record a movie of 100 or more frames a second. It is used for observing cavitation, crash test and so on, because it can record fast movement without destruction and contact. Since the age of film cameras, a lot of machines for high-speed shooting have been invented. In these days, development of high-quality sensors and high-capacity memories make it possible to produce high-speed camera which is able to record 1,000,000 or more fps and to produce a high-speed camera for consumer use. In this study, high-speed cameras are used for recording sounds. When recording sound using high-speed cameras, a frame rate of camera corresponds to a sampling rate of A/D converter. From sampling theorem, a high-speed camera can record sounds, which include lower frequency components than half frequency of its frame rate. Considering the performance of a high-speed camera, auditory sounds can be fully recorded by the high-speed camera.

Image Processing for Obtaining Vibration

Object motions include sound vibration are obtained from high-speed video. There are some methods for objects tracking. At this time, vibrations were tracked using a center of gravity of the image. We assume the change of a position of the center of gravity as object movement, because each pixel value changes with the object movement in the video. To obtain a sound vibration from high-speed video, firstly, analyzed region is set on the image. Secondly, calculate a position of the center of gravity of the region on each frame.

In the set region (H px×W px), a position of the center of gravity

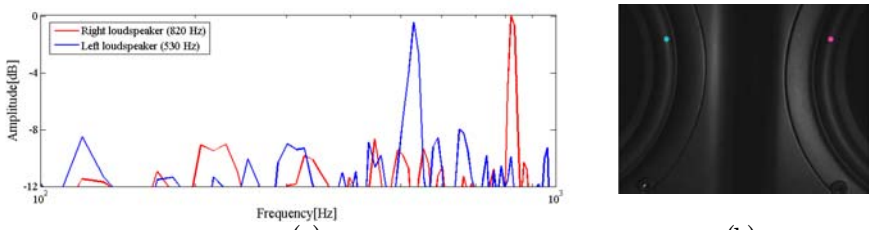
$$\begin{bmatrix} G_i \\ G_j \end{bmatrix} = \frac{1}{M} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} g[i][j] \begin{bmatrix} i \\ j \end{bmatrix}, \tag{1}$$

where M is total pixel values of the region, (i, j) is coordinate of pixel and g[i][j] is pixel value. And then, the changes of the position are obtained.

Recording Sounds of Loudspeakers

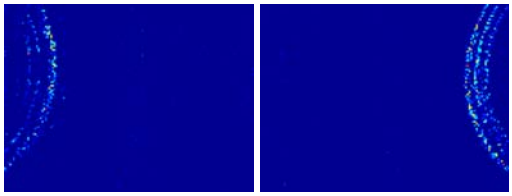
To confirm our method using high-speed camera for recording sound, sounds and locations of loudspeaker were obtained. Two loudspeakers were set side by side. The loudspeaker on the left was driven by 530 Hz sinusoidal sound and the right one was driven by 820 Hz sinusoidal sound. The high-speed camera was MEMRECAM HX-3 (nac image technology) and shot at 6,000 fps. Figure 1 shows original image we shot. Figure 2 shows a spectrum at the red and blue point illustrated in Fig. 2 (b). Frequency spectrums on each 10×10px region were attained from change of center of gravities. Figure 3 (a) shows 500-550 Hz vibration areas and Fig. 3 (b) shows 800-850 Hz vibration area. In the case of using a high-speed camera for recording sounds, it is possible to separate each sound from high-speed videos, because a high-speed camera records every point at the same time within the range of the camera. From Fig. 3, sound vibrations of loudspeakers were separated completely.

FIGURE 1. Loudspeakers image recorded using high-speed camera.
The left one was driven by 530 Hz sinusoidal sound and the right one was driven by 820 Hz sinusoidal sound.



(a) (b)

FIGURE 2. Frequency spectrum obtained from high-speed video of two loudspeakers.
Blue and red lines show the frequency spectrum at the point colored with blue and red in (b).



(a) (b)

FIGURE 3. Analyzed images of loudspeakers.
(a) Analyzed loudspeaker image. 500-550 Hz vibration areas are colored.
(b) Analyzed loudspeaker image. 800-850 Hz vibration areas are colored.

EXTRACTION OF VOICE FROM HIGH-SPEED VIDEO

Sound Propagation Through the Soft Biological Tissue

A voice source generated in vocal cords, resonant in vocal tract to be voice. The voice is radiated not only from a mouth and a nostril but also through soft biological tissues [11]. In this section, we discuss about acoustic damping through the soft biological tissues. Acoustic behaviors in the soft biological tissues have been mainly studied in the field of ultrasonic waves and it is known that sound velocity in the soft biological tissue is 1540 m/s. When the sounds propagate through the soft biological tissue, sounds damp in proportion to the frequency to the n^{th} ($1<n<2$) power. The damping constant

$$\alpha = \alpha_0 f^n, \tag{2}$$

where α_0 is constant of proportion [12]. At this time, n=1.2, α_0=0.5 dB/cm/MHz.

The cheek thickness average of Japanese man is 7.8 mm [13], and sounds of 1,000 Hz decay 1.55×10^{-3} dB inside the cheek from the formula. If the incident on the soft biological tissue is considered, the sound level is more decay than this calculation.

Acoustic damping was examined using an oscillatory actuator. The actuator (RION BR-41) vibrated in 450 Hz, and a cheek is contacted with the actuator. The vibrations were measured using the laser Doppler vibrometer (Polytec OFV-505) as shown in Fig. 4. To enhance the measurement accuracy, reflection tapes were put on the surface of the actuator and the cheek skin. Figure 5 shows the actuator and Fig. 6 shows the velocity of the actuator and the cheek. The vibration decreases about 25% through the cheek.

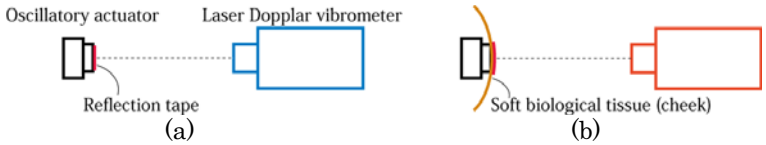


FIGURE 4. Experimental settings
(a) Measuring an oscillatory actuator.
(b) Measuring a cheek vibrated by an oscillatory actuator.

FIGURE 5. Used oscillatory actuator (RION BR-41)

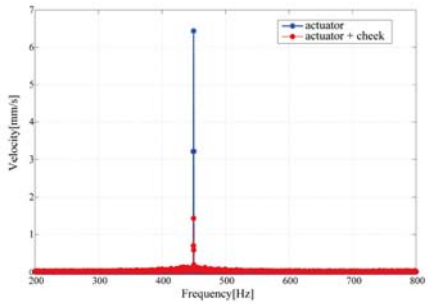


FIGURE 6. Velocity of the actuator and actuated cheek. It was measured using the laser Doppler vibrometer.

Extraction of Vowel Sound

We recorded two subjects saying /a/ or /tawara/ using two high-speed cameras. In addition to high-speed cameras, voices and the vibration of cervical part were recorded using microphones and the laser Doppler vibrometer (Polytec OFV-505). For the purpose of enhancement of the measurement accuracy, a reflection tape was put on the cervical part. Subjects used the eye mask to save their eyes though level of the laser is not dangerous.

Voice information were extracted from high-speed videos of lips and cervical part of the examinee saying /a/. As illustrated in Fig. 7, two high-speed cameras (Photron FASTCAM SA-X, 10,000 fps) and the laser Doppler vibrometer (Polytec OFV-505) were used in this experiment. Figure 8 shows recorded image. A rectangle indicated on the Fig. 8 is the analyzed region, which contains edge of cervical part.

The sound spectrum recorded using the microphones and the vibration spectrum of cervical part recorded using the laser Doppler vibrometer are shown in Fig. 9 and Fig. 10, respectively. It can be seen from Fig. 9 that the sound recorded using the microphone includes the filtering effect of vocal tract in addition to the vocal cord vibration. From Fig. 10, vibration of cervical part recorded using the laser Doppler vibrometer include the vocal cord vibration. The vibration spectrum recorded using the high-speed camera is showed in Fig. 11. Figure 11 has several frequency components in common with Fig. 9 and Fig. 10. From the results, it is possible that voices are extracted from high-speed videos of the cervical part. We find 100 Hz on Fig. 11 and it is thought to be caused by cameras vibration causes by fan or lights.

A method with multiple analyze regions were attempted to reduce noise. Five analyzed regions are illustrated in Fig. 12. The average vibration spectrum of five regions is showed in Fig. 13. Comparing Fig. 13 with Fig. 11, it is obvious that noises are reduced in Fig. 13.

Lips vibrations were also analyzed. Six illustrated rectangles (30 px ×30 px) in Fig. 14 are analyzed regions, which contains a edge between the upper lip and the oral cavity. The longitudinal vibration spectrum of lips is showed in Fig. 15. The 150 Hz component appears in Fig. 15, but other frequency components in voice are not founded.

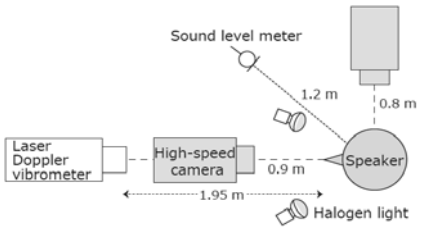


FIGURE 7. Experimental settings examinee saying /a/.

FIGURE 8. Recorded image of cervical part with analyze region.

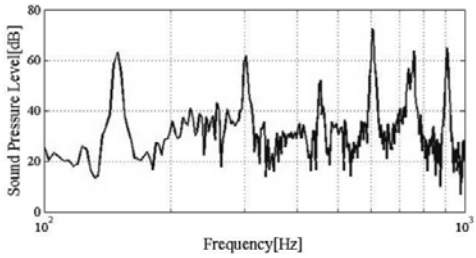


FIGURE 9. Sound spectrum of /a/ recorded using microphone.

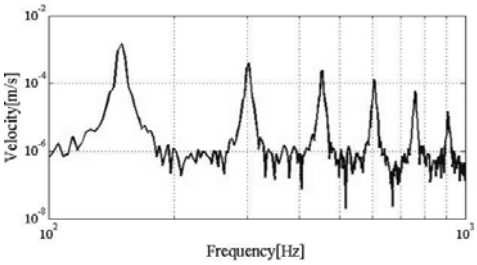


FIGURE 10. Cervical part vibration spectrum of /a/ recorded using laser Doppler vibrometer.

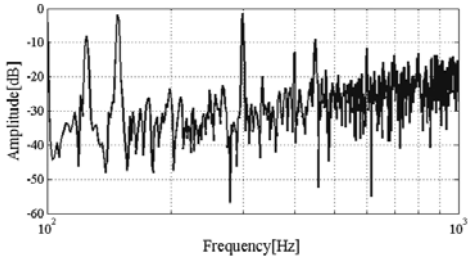


FIGURE 11. Cervical part vibration spectrum of /a/ recorded using high-speed camera with a analyze region.

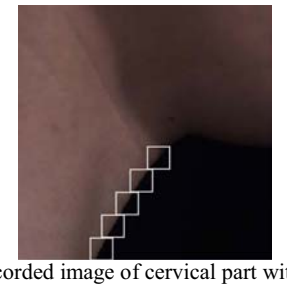


FIGURE 12. Recorded image of cervical part with 5 analyze regions.

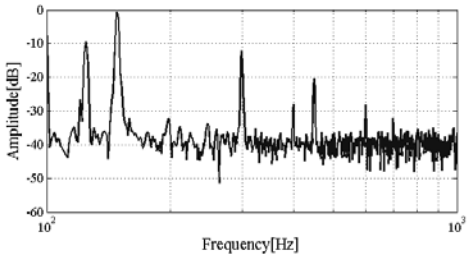


FIGURE 13. Cervical part vibration spectrum of /a/ recorded using high-speed camera with 5 analyze regions.

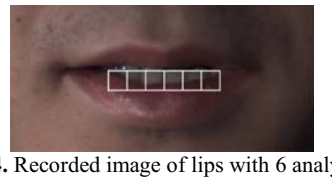


FIGURE 14. Recorded image of lips with 6 analyze region.

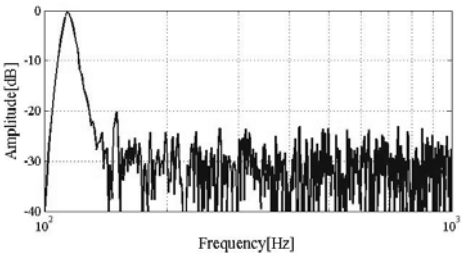


FIGURE 15. Lips vibration spectrum of /a/ recorded using high-speed camera with 6 analyze region.

Extraction of Words Voice

Voice information were extracted from high-speed video of cervical part of the examinee saying /tawara/. The high-speed camera (nac imaging technology MEMRICAM HX-3, 10,000 fps) and the laser Doppler vibrometer (Polytec OFV-505) were set as Fig. 16. At this time, the light was set opposite to camera to record the cervical part more clearly. Recorded image and analyzed regions are showed in Fig. 17. The spectrogram of sound recorded with the microphone is shown in Fig. 18, the cervical vibration recorded using the laser Doppler is shown in Fig. 19 and the average cervical vibration recorded using the high-speed camera is shown in Fig. 20. Each waveform is shown under each spectrogram. From these Figures, changes of voice, like intonation, are appeared in the spectrogram with high-speed camera.

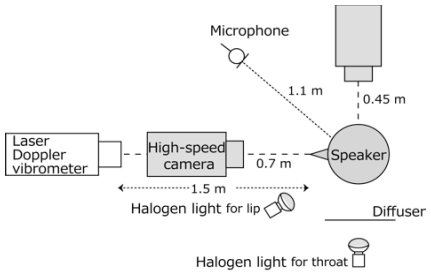


FIGURE 16. Experimental settings examinee saying /tawara/.

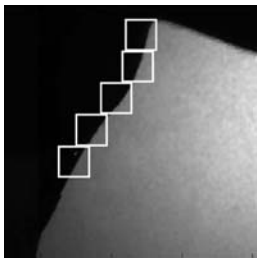


FIGURE 17. Recorded image of cervical part with 5 analyze regions.

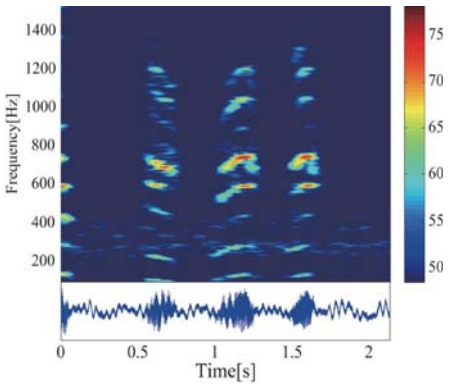


FIGURE 18. Sound spectrogram and waveform of /tawara/ recorded using microphone.

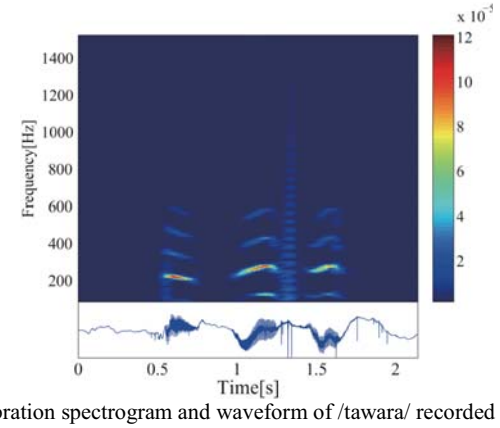


FIGURE 19. Cervical part vibration spectrogram and waveform of /tawara/ recorded using laser Doppler vibrometer.

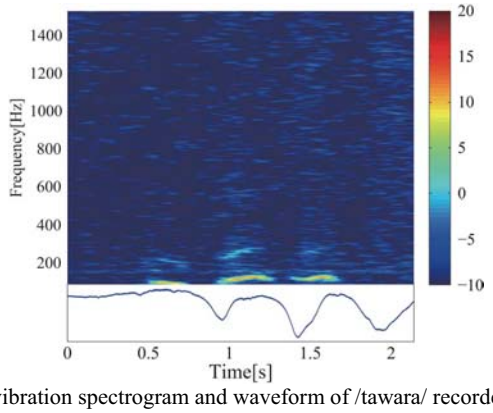


FIGURE 20. Cervical part vibration spectrogram and waveform of /tawara/ recorded using high-speed camera.

CHANGE OF LIP SHAPE DEPENDS ON WORDS

Lips shapes when speaker saying Japanese vowel /a/, /i/, /u/, /e/, /o/ were observed. Considering lips videos will be taken from side at the same time as cervical part taken, we examined changes of lips shapes on side face. Recorded images are shown in Fig.21. We extract a contour of front face and set local maximum values in a direction to front face as 4 feature points, which are nose, upper lip, bottom lip and chin as showed as Fig. 22.

Figure 23 shows positional relationships of each feature among 5 vowels. In this figure, normalizing the distance and angle between nose point and chin point comparison. Distance ratios between features are shown in Fig. 24. These results indicate distinctions between vowels well. Figure 23 indicates that lips are located backward since they are extended transversely when uttering /i/ and /e/. And lips are located forward than face when uttering /o/. Meanwhile, when speaking /a/ and /u/, lips are located near the nose-chin line. From Fig. 24, it is clear that the distance from upper lip to bottom lip on the situation of /e/ and /a/ is longer than /i/ and /u/, respectively. Considering these results, vowels can be estimated from lips distance from nose-chin line and distance ratios between features.

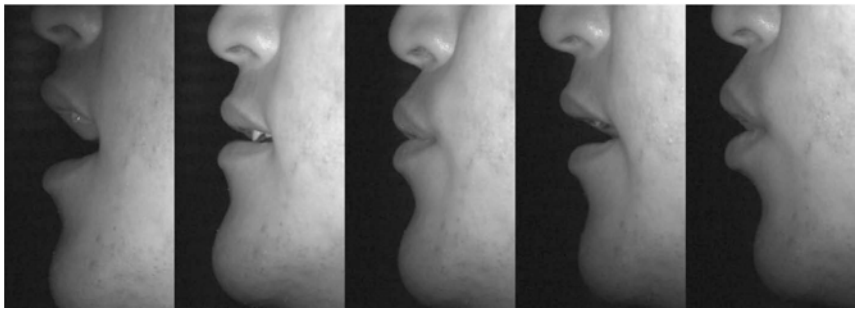


FIGURE 21. Recorded lips shapes of 5 vowels from side.

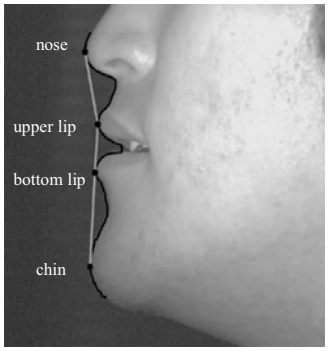


FIGURE 22. Defined 4 features, nose, upper lip, bottom lip and chin.

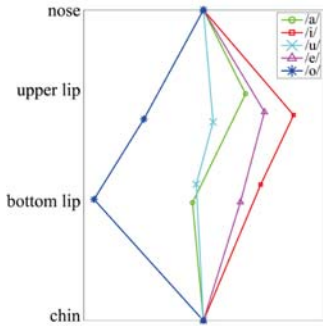


FIGURE 23. Positional relationships of upper lip and bottom lip for nose and chin.

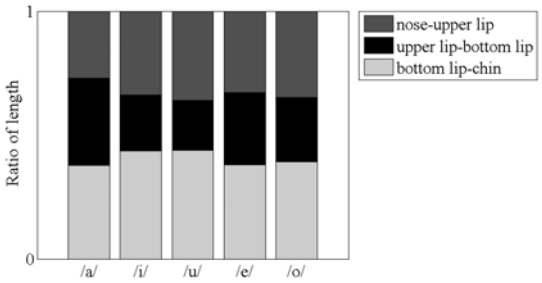


FIGURE 24. Distance ratios between 4 features.

CONCLUSIONS

The purpose of this study is to extract speech voice information from high-speed videos. The results show that the voice information, includes intonation are possible to be extracted from high-speed videos of cervical parts. However, it was hard to extract voice information from vibrations of lips. Regarding shape of lips, we set 4 feature points to distinguish vowels.

For the reconstruction of speech voice, it is necessary to consider improving S/N, and shooting methods or region. Furthermore, giving a boost to extraction of accurate voice and consonant are subjects to our future studies.

REFERENCES

1. Y. Oikawa, M. Goto, Y. Ikeda, T. Takizawa and Y. Yamasaki, “Sound Field Measurement Based on Reconstruction from Laser Projections, “ *Proc. ICASSP*, IV, pp. 661-664, (2005).
2. Y. Oikawa, T. Hasegawa, Y. Ouchi, Y. Yamasaki and Y. Ikeda, “Visualization of soundfield and sound source vibration using laser measurement method,” *Proc. ICA* 2010, 898, (2010.8).
3. T. Sakoda and Y. Sonoda, “Visualization of sound field with uniform phase distribution using laser beam microphone coupled with computerized tomography method,” Acoust. Sci. & Tec., Vol. 29, No. 4, pp.295-299, (2008).
4. T. Kanemoto, R. kogure, T. mikoshiba, O. Yasojima, Y. Oikawa and Y. Yamasaki, “Measurement of Sound Field with Optical Wave Microphone Using Refraction and Diffraction,” *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 709-710, (2012)(in Japanese).
5. J. Sakai, S. Takeoka, Y. Oikawa and Y. Yamasaki, “The 3-DAnalysis of vibrating objects utilizing a high speed camera,” *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 1493-1494, (2008)(in Japanese).
6. S. Takeoka, O. Yasojima, J. Sakai, Y. Oikawa and Y. Yamasaki, “Measurement of particle velocity distribution by PIV method with high speed camera,” *Tech. Rep. ASJ*, AI2010-3-03, (2010.10) (in Japanese).
7. M. Akutsu, Y. Oikawa, “Extraction of sound field information from flowing dust captured with high-speed camera,” *Proc. ICASSP*, pp.545-548, (2012).
8. M. Akutsu, Y. Oikawa and Y. Yamasaki, “Extraction of sound field information from high-speed movie of flowing dust,” Acoust. Sci. & Tec., Vol. 33, No. 5, pp.316-319, (2012).
9. K. Mase and A. Pentland, “Automatic Lipreading by Optical-flow Analysis,” *Tech. Rep. ITEJ*, 13(44), pp.7-12, (1989)(in Japanese).
10. R. Kaucic, B. Dalton and A. Blake, “Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications,” *Proc. European Conf. Computer Vision*, pp. 376-387, (1996).
11. T. Kitamura, “Measurement of vibration of skin during phonation using scanning vibrometer,” *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 329–330, (2012)(in Japanese).
12. I. Akiyama, “On The Effects of Frequency Dependent Attenuation of Biological Tissues in Ultrasonic Imaging – Transmitting Waveform Relevant to Image Depth-,” *Tech. Rep. IEICE*, EA2011-111, pp. 43-48, (2012)(in Japanese).
13. N. Mori, M. Terasima, K. Tokumori, A. Nakajima, Y.Aoki, S. Hashimoto, “Head bioinstrumentation and construction of face standard three-dimensional physical model of modern Japanese adults of both sexes using three-dimensional CT images,” *J. Anthropological Soc. Jpn.*, 111(1), pp35-49, (2003).