Journal of Information Processing Vol.25 331–340 (Apr. 2017)

[DOI: 10.2197/ipsjjip.25.331]

Recommended Paper

Gesture Recognition Method Utilizing Ultrasonic Active Acoustic Sensing

Hiroki Watanabe[1,2,a)] Tsutomu Terada[1,3,b)] Masahiko Tsukamoto[1,c)]

Received: May 7, 2016, Accepted: January 10, 2017

Abstract: We propose a method for gesture recognition that utilizes active acoustic sensing, which transmits acoustic signals to a target, and recognizes the target's state by analyzing the response. In this study, the user wore a contact speaker that transmitted ultrasonic sweep signals to the user's body and a contact microphone that detected the ultrasound propagated through the body. The propagation characteristics of the ultrasound changed depending on the user's movements. We utilized these changes to recognize the user's gestures. One of the important novelty features of our method is that the user's gestures can be acquired not only from the physical movement but also from the user's internal state, such as muscle activity, since ultrasound is transmitted via both the user's internal body and body surface. Moreover, our method is not adversely affected by audible-range sounds generated by the environment and body movements because we utilize ultrasound. We implemented a device that uses active acoustic sensing to effectively transmit/detect the ultrasound to/from the body and investigated the performance of the proposed method in 21 contexts with 10 subjects. The evaluation results confirmed that the precision and recall are 93.1% and 91.6%, respectively when we set 10% of the data as training data and the rest as testing data in the same data set. When we used the data set for training and the other data set for testing in the same day, the precision and recall are 51.6% and 51.3%, respectively.

Keywords: wearable computing, gesture recognition, ultrasound, active acoustic sensing

1. Introduction

A gesture recognition method is an important technique these days for various purposes including developing intuitive interfaces. Commercial devices, such as Kinect [4] and Leap Motion [5], are already available. In the wearable computing environment, they are utilized for the life log and hands-free interface by using wearable sensors. A lot of gesture recognition methods have already been studied, and typical sensors for recognizing user gestures are accelerometers [18] and cameras [22]. However, when using the former, it is difficult to acquire the internal/surface state of a body. Use of the latter is effective only in the range where the camera can get an image. Thus, it is not suitable for the wearable computing environment. Recognizing aspects of the user's internal state, such as muscle activity, is expected to enrich the user's experience in wearable computing. Although electromyography (EMG) sensors [8] are typically used to recognize the aspects of the internal state of the body, the changes in the EMG sensor values are small, and EMG sensors are adversely affected by electrical noise. Thus, we consider that it is difficult to utilize EMG sensors in daily life.

In this study, we focused on the active acoustic sensing method, which transmits acoustic signals to a target and recognizes the target state by the response. We applied this technique to a human body. The user wore a contact speaker that transmitted ultrasound to his/her body and a contact microphone that detected the ultrasound that propagates through his/her body. The propagation characteristics of the ultrasound changed depending on the aspects of the user's state, such as gestures and muscle activities. We utilized these properties to recognize the user's gestures. One of the important novelty features of our method is that the user's gestures can be acquired not only from the apparent movement, such as limb movement (physical movement) but also from the user's internal state, such as muscle activity, since ultrasound is transmitted via both the user's internal body and body surface. In the conventional method, the combination use of multiple sensors, such as accelerometers and EMG sensors, is required to acquire these contexts. Our method can acquire these contexts by using only a contact microphone and a contact speaker, which are simple and cheap. Moreover, since we utilized ultrasonic-range sound, our method is not adversely affected by audible-range sounds generated by the environment and body movements. We implemented a prototype device and investigated the performance of the proposed method in 21 contexts with 10 subjects. The evaluation results confirmed that the precision and recall are 93.1% and 91.6%, respectively when we set 10% of the data as training data and the rest as testing data in the same data

[1] Graduate School of Engineering, Kobe University, Kobe, Hyogo 657–8501, Japan
[2] Research Fellow of Japan Society for the Promotion of Scienece, Chiyoda, Tokyo 102–0083, Japan
[3] PRESTO, Japan Science and Technology Agency, Chiyoda, Tokyo 102–0076, Japan
[a)] hiroki.watanabe@stu.kobe-u.ac.jp
[b)] tsutomu@eedept.kobe-u.ac.jp
[c)] tuka@kobe-u.ac.jp

The preliminary version of this paper was published at Multimedia, Distributed, Cooperative, and Mobile Symposium (DICOMO 2015), July 2015. The paper was recommended to be submitted to Journal of Information Processing (JIP) by the chief examiner of SIGUBI.

set. When we used the data set for training and the other data set for testing in the same day, the precision and recall are 51.6% and 51.3%, respectively.

The contributions of this paper are as follows:

1) We proposed a gesture recognition method utilizing the propagation characteristics of the ultrasonic sweep signals that are transmitted to the internal body and the surface of the body.
2) We designed and implemented the microphone and the speaker to effectively transmit/detect sound to/from the body. Our device reduces the deviation in the device wearing position by utilizing suspension and adhesive force.
3) We investigated the performance of the proposed method assuming daily life use considering re-attaching the device for another day.

This paper is organized as follows. In Section 2, we describe related work. In Section 3, the proposed method is presented. The implementation is described in Section 4. The recognition rate of our method is discussed in Section 5. Finally, Section 6 concludes our research.

2. Related Work

2.1 Gesture Recognition Method

There are a lot of gesture recognition methods, and accelerometers are generally utilized. Murao et al. [18] investigated the effects on recognition accuracy of changing the number and positions of sensors and the number and kinds of gestures by using a test mobile device with nine accelerometers and nine gyroscopes. Watanabe et al. [26] developed an activity and context recognition method where the user carries a neck-worn receiver comprising a microphone and small speakers on his/her wrists that generate ultrasounds. Stiefmeier et al. [24] presented a method for continuous activity recognition based on ultrasonic hand tracking and motion sensors attached to the user's arms. Gupta et al. [13] presented SoundWave, a technique that leverages the speaker and microphone already embedded in most commodity devices to sense in-air gestures around the device. Kinect [4] can recognize the user's gestures by utilizing an RGB camera and depth sensor. Pirkl et al. [20] described the design and implementation of a cheap, low power, and easily wearable system for tracking the relative position and orientation of body parts by utilizing magnetic field technology. Fukui et al. [12] proposed an approach to hand shape recognition based on wrist contour measurement. Sato et al. [23] proposed *Swept Frequency Capacitive Sensing* technique that can not only detect a touch event but also recognize complex configurations of the human hands and body.

In these studies, although physical movements that are visually recognized from the outside are recognized precisely, it is difficult to recognize aspects of the internal state of the body such as muscle activity. Moreover, some methods are adversely affected by environment, for example, it is difficult to use infrared sensors at the outdoors because sunlight includes infrared and causes misrecognition.

Mokaya et al. [16] developed a wearable system for determining muscle activation in high motion exercise scenarios. However, this study is different from our method in terms of using accelerometers. EMG sensors can recognize muscle activities. Amma et al. [8] presented their results on using EMG sensor arrays for finger gesture recognition. However, changes in the EMG sensor values are small, and EMG sensors are adversely affected by electrical noise. Thus, we consider that it is difficult to utilize EMG sensors in daily life.

2.2 Active Acoustic Sensing for Objects

Active acoustic sensing is a method for estimating the state of objects by transmitting acoustic signals and analyzing the response. In the industrial field, this technique is known as a non-destructive testing technique [6]. Surface acoustic wave touch screens also utilize acoustic signals to detect the user's touch position [9].

In research, Ono et al. [19] presented an acoustic touch sensing technique called *Touch & Activate*. It recognizes a rich context of touches including grasping existing objects by attaching only a vibration speaker and a piezoelectric microphone paired as a sensor. Laput et al. [14] developed *Acoustruments*, which are low-cost, passive, and power-less mechanisms made from plastic, that can bring rich, tangible functionality to handheld devices. They attach structural elements to a handheld device along the speaker-microphone pathway to characteristically alter the acoustic output. SoQr [11] is a sensor that can be attached to an external surface of a household item to estimate the amount of content inside it. The sensor consists of a speaker and a microphone.

In these studies, they apply active acoustic sensing to objects. We apply this technique to the human body and recognize the user's gestures.

2.3 Active Acoustic Sensing for Human Body

Active acoustic sensing is also applied to the human body. In the medical field, active acoustic sensing is widely used to see internal body structures such as tendons, muscles, and internal organs [10]. In research, Mujibiya et al. proposed a sensing technique based on transdermal low-frequency ultrasound propagation [17]. This technique enables pressure-aware continuous touch sensing as well as arm-grasping hand gestures. Takemura et al. proposed a wearable sensor system that measures the angle of an elbow and the position of a tapping finger using bone-conducted sound [25]. It consists of two microphones and a speaker, and they are attached to the forearm. In these studies, they utilized certain fixed frequencies. Moreover, they do not detect daily gestures with large limb movements, such as those of walking and jogging.

In this study, we utilized a sweep signal, which is a signal where the frequency increases/decreases with time. Moreover, we transmitted ultrasound to the human body and utilized the reflected ultrasonic wave from the internal body and the surface wave on the body whereas previous methods [17] utilized only the surface wave on the body. Additionally, we considered daily gestures including walking and jogging.

3. Proposed Method

In our proposed method, the user carried wearable speaker and microphone on his/her body, as shown in **Fig. 1**. The contact

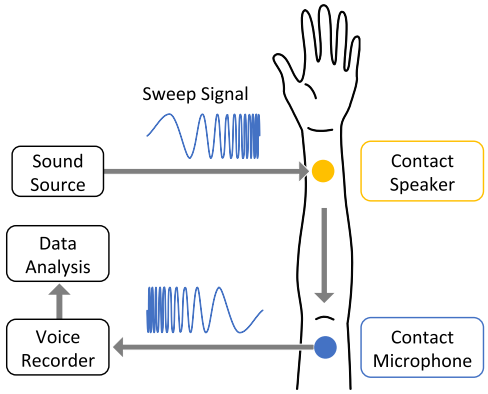

Fig. 1 System configuration.

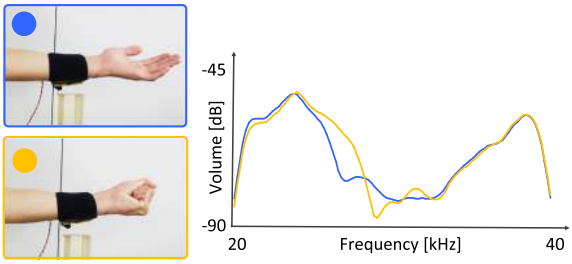

Fig. 2 Changes in frequency spectrum.

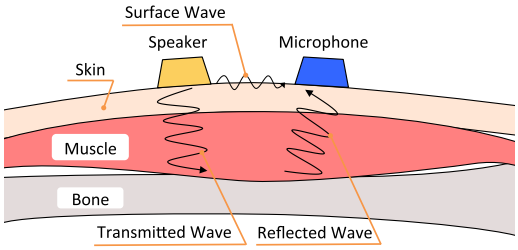

Fig. 3 Propagation of ultrasound.

speaker repeatedly transmitted ultrasonic sweep signals to the human body, and the contact microphone detected the ultrasound that propagates through the body. We performed fast Fourier transform (FFT) on the input from the contact microphone. In this study, the sampling rate was 96 kHz, and the number of FFT samples was 8,192. As shown in **Fig. 2**, the frequency spectrum was changed by the user's state, for example, during postures and muscle activities. We utilized these features to recognize user gestures. Since the sound from the contact speaker was ultrasound, humans could not hear the sound propagated in the air.

3.1 Ultrasound Propagation in Human Body

Figure 3 shows the propagation of ultrasound. The signal from the sound source vibrates the contact speaker, and the contact speaker vibrates the skin. A portion of the sound wave is reflected on the surface (surface wave), and the other portion propagates through the body (transmitted wave) [15]. When the sound wave crosses a boundary between two media, a portion of the energy is reflected (reflected wave) and a portion continues in a relatively straight line. This reflectance is caused by a difference in the acoustic impedance between the two media. The larger the difference, the more sound waves are reflected. The human body consists of skin, fat, blood, bone, and so on. These components have their own acoustic impedance. Sound waves transmit and reflect on these tissue boundaries and have unique propagation characteristics. The body movements affect the positional relationship of these components and change their shape. These changes of components affect the ultrasound propagation in the human body and dynamically change the ultrasound propagation characteristics. We utilized these changes in frequency to recognize user gestures. In previous study [17], they utilized commodity ultrasonic transducer and receiver. In this study, we designed the contact speaker/microphone considering the acoustic impedance between device and skin to effectively transmit/detect ultrasound to/from the human body, and we utilized the reflected wave from the internal parts of the body and the surface wave on the body.

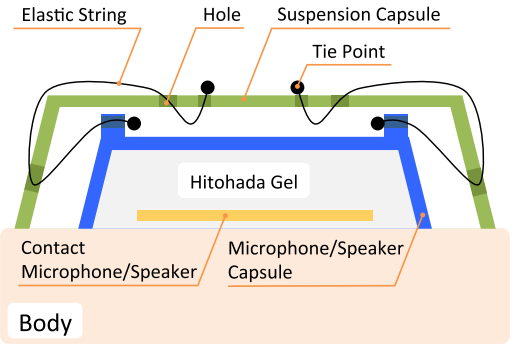

Fig. 4 Microphone/speaker configuration.

Transmitting ultrasound to the human body is widely used in the medical field. It is basically harmless to the human body as long as the ultrasound intensity is below a certain threshold. We follow the guideline that Canada, Japan, Russia, and the International Radiation Protection Agency recommends a maximum level of 110 dB for safe operation for frequencies from 25 to 50 kHz [2].

3.2 Microphone/Speaker Design

To effectively transmit/detect ultrasound to/from the human body, we have to consider the difference in acoustic impedances between the contact microphone/speaker (piezoelectric sensor) and the skin. The acoustic impedance of typical ceramic piezoelectric sensors and the human body are $30–35 \times 10^6$ kg/m^{2}s and approximately 1.5×10^6 kg/m^{2}s, respectively. Therefore, a major portion of the ultrasound is reflected on the skin by just wearing the contact microphone/speaker. BodyBeat [21] solved the problem of acoustic impedance difference by detecting the sound from the body via a silicon diaphragm. Based on this idea, we designed the microphone/speaker for active acoustic sensing of the body, as shown in **Fig. 4**. The 3D printed capsule was filled with Hitohada gel (hardness: 15) [3]. A piezoelectric sensor was embedded in the hardened gel. Hitohada gel is a super soft resin for modeling, which has an acoustic impedance of approximately 1.7×10^6 kg/m^{2}s. The difference in acoustic impedance between the body and the gel is much less than that between the body and the ceramic piezoelectric sensor. We compared the conventional piezoelectric sensors with the proposed device on the human body. We attached the device to the wrist, and the speaker transmitted the ultrasonic sweep signal and the microphone de-

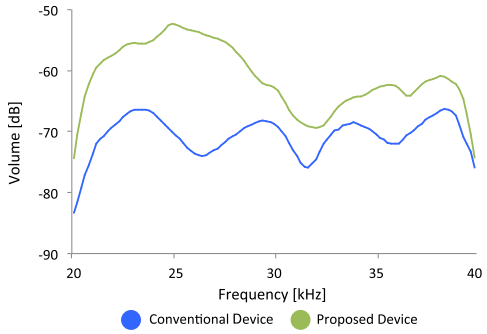

Fig. 5 Comparison of the conventional device and the proposed device.

tected the ultrasound propagated in the body. **Figure 5** shows an example of comparing the frequency spectrum for both devices. As shown in this figure, the volume of ultrasound by using the proposed device was larger at each frequency than that by using the conventional device because the proposed device can more effectively transmit/detect ultrasound than the conventional device. In this study, the detected wave consisted of surface wave and reflected wave. A future work will be to separate these two waves and utilize each characteristic.

We also considered that deviation in the device wearing position and friction noise that includes ultrasound caused by user movement affect the recognition result. Thus, based on the method in Ref. [21], we designed a microphone/speaker attachment that reduces the deviation in the device position caused by the user's movements as shown in Fig. 4. The microphone/speaker is attached to the suspension capsule with four elastic strings. The suspension allows for approximately four millimeters of all sides and vertical movements. This helps to keep the microphone/speaker in the same position because the deviation caused by the body movements and external impacts only affect the suspension capsule and the microphone/speaker can keep a certain position. The deviated suspension capsule returns to the original position by the tension of the elastic strings so as not to strike the internal microphone/speaker capsule. Moreover, Hitohada gel has an adhesive force. The microphone/speaker sticks to the skin and keeps its original position. Although the adhesive force became weaker after we had used it many times, wiping the gel with water or alcohol revived it. In this study, the designed suspension capsule is 38 millimeters in diameter. It is relatively big for daily use. However, we plan to miniaturize the device in the same configuration in the future.

3.3 Sweep Signal

We adopted the ultrasonic sweep signal as an acoustic signal from the contact speaker. A sweep signal is a signal whose frequency increases/decreases with time. Since it includes various frequencies, we can acquire more features than a fixed frequency. In this study, we used a sweep signal shifting from 20 to 40 kHz over 20 milliseconds and a contact speaker that transmits the sweep signal repeatedly. Since the sound transmitted from the speaker is ultrasonic range sound (the sound of more than 20 kHz), it does not annoy the user and the surrounding people. Moreover, 40 kHz has enough margins for the Nyquist frequency of high-end mobile phones and audio players. Additionally, when we analyze the acquired data, the window for calculating FFT (approximately 85.3 milliseconds) includes at least four cycles of the sweep signal. This ensures that we do not have to consider the timing of the sweep signal when analyzing the acquired data, and we can acquire a stable frequency response of four cycles of sweep signals.

3.4 Gesture Recognition Method

We calculated FFT for the input of the contact microphone and obtained the frequency spectrum. The sampling rate was 96 kHz. The number of samples for calculating FFT was 8,192 without any overlaps. We focused on only the 20 to 40 kHz range, which is the transmitted sweep signal range, and extracted features from 20 to 40 kHz of the frequency spectrum. We extacted 25 features as follows.

- Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs are commonly used for audio and speech recognition [27]. Note that we did not use a mel filter bank when we calculated MFCCs because we did not have to consider human hearing characteristics and do need equally observe the target range. Therefore, we utilized 20 triangular overlapping windows that have equal intervals on the hertz scale instead of the mel filter bank.
- Spectral Centroid: the spectral centroid is the balancing point of the spectral power distribution. It is calulated by the following formula. $f(k)$ represents the frequency magnitude of bin number k in the range of 20 to 40 kHz.

$$\text{Spectral Centroid} = \frac{\sum_{k=1}^{N} kf(k)}{\sum_{k=1}^{N} f(k)} \tag{1}$$

- Spectral Flux: the spectral flux is the spectral amplitude difference of two adjacent frames. $f_t(k)$ and $f_{t-1}(k)$ represent the current frame and the last frame, respectively.

$$\text{Spectral Flux} = \sum_{k=1}^{N} (f_t(k) - f_{t-1}(k))^2 \tag{2}$$

- Spectral Skewness: the skewness is a measure of the asymmetry of the data around the sample mean. μ and σ represents mean and standard deviation, respectively.

$$\text{Spectral Skewness} = \frac{\sum_{k=1}^{N} (f(k) - \mu)^3}{N\sigma^3} \tag{3}$$

- Spectral Kurtosis: the kurtosis is a measure of how outlier-prone a distribution is.

$$\text{Spectral Kurtosis} = \frac{\sum_{k=1}^{N} (f(k) - \mu)^4}{N\sigma^4} \tag{4}$$

- Spectral Rolloff of 93%: it is defined as the frequency bin below which 93% of the distribution is concentrated.

$$\text{Spectral Rolloff} = max\left(h \,\middle|\, \sum_{k=1}^{h} f(k) < threshold\right) \tag{5}$$

Moreover, we calculated the mean and variance of 25 features for detecting the changes in the time series. Thus, we acquired 50 features in total. The window size for calculating the mean and variance was 30 (approximately three seconds) because

some contexts used in evaluation were approximately three seconds per action. We utilized these features to recognize gestures. We utilized WEKA [7] for classification and selected IB1 for the classifier, which is the nearest neighbor method implemented on WEKA. Although other classifiers, such as support vector machine and decision tree, can be considered, we selected IB1 because the recognition result was the best.

4. Implementation

We implemented a prototype device. **Figure 6** shows the implemented microphone and speaker. The contact microphone was a Murata 7BB-20-6L0, and the contact speaker was a Thrive OMR20F10-BP310. These piezoelectric sensors were embedded into a 3D printed ABS capsule that was filled with Hitohada gel (hardness: 15). The microphone/speaker was attached to the suspension capsule with four elastic strings and was attached to the supporter by a snap button. Thus, these devices could easily be attached/detached from the supporter. These devices were fixed by the supporter, as shown in the bottom right of Fig. 6.

Figure 7 shows the whole prototype device. The ultrasound transmitting device consists of an ARM mbed NXP LPC1768, lithium-ion battery (850 mAh, 3.7 V), boosting module (LMR62421), and power cell LiPo charger/booster. This device can be attached to the outside of the supporter by velcro tape. A sweep signal sound file (which shifts from 20 to 40 kHz over 20 milliseconds) is located in mbed local storage. It reads the sound file and outputs from AnalogOut. The sweep signal was made by the chirp (sweep) signal function of Audacity, which is a free software for recording and editing sounds [1]. To smoothly repeat the sound, we processed the one millisecond fade in and fade out for the beginning and end of the sound file, respectively. The output signal from mbed is boosted to 12 V by the boosting module to sufficiently vibrate the contact speaker. Since the output volume of the ultrasound was at most 60 dB, our signal was below the safety threshold of 110 dB [2].

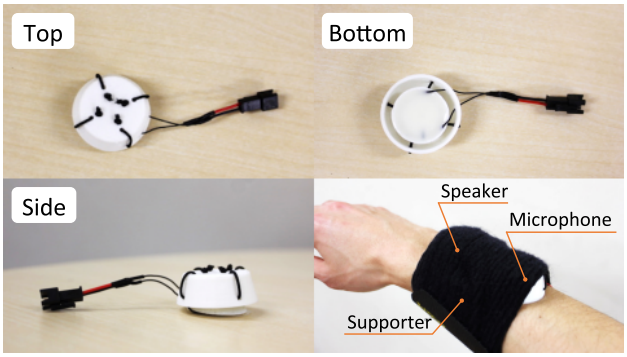

Fig. 6 Implemented microphone and speaker.

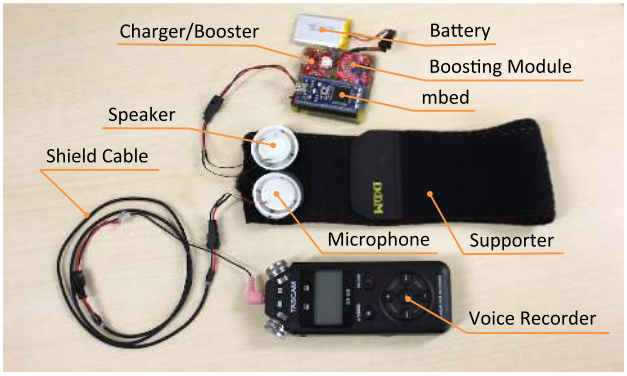

Fig. 7 Prototype device.

We utilized a voice recorder TASCAM DR-05 to record the data. The recording sampling rate was 96 kHz, and the quantization bit was 16 bit. The voice recorder records the signal from the microphone via shield cable. The signal from the microphone was so small that we used shield cable to reduce the noise, and the signal was amplified in the voice recorder.

We used an Apple MacBook Pro (CPU: Intel Core i7 3.1 GHz, RAM: 16 GB) to analyze the data. We implemented software for data analysis by using C++.

5. Evaluation

5.1 Wearing Position

We evaluated the system to determine the appropriate wearing position of the device. Although our purpose is recognizing gestures, we investigated how we can classify the state of the body depending on the wearing position. The subject wore the device in 12 positions, as shown in **Fig. 8**. We assumed four statuses at each position: the angle of the upper arm and the forearm was 180 degrees, the angle of the upper arm and the forearm was 180 degrees with tightening muscles, the angle of the upper arm and the forearm was 90 degrees, and the angle of the upper arm and the forearm was 90 degrees with tightening muscles. The same was applied to the leg between the thigh and calf. We recorded the data for 30 seconds at each status and extracted the features described above. We set 10% of the features as training data, and the rest was testing data. The three subjects were 22 to 26-year-old males.

Table 1 shows the average results of the evaluation. The recognition rate was relatively high in all positions except position 7 and 10. In these two positions, the distance between the microphone and the speaker was the longest in all wearing positions. Therefore, the ultrasound attenuated during propagation, and it was difficult to recognize precisely.

Therefore, from the point of view of wearability and recog-

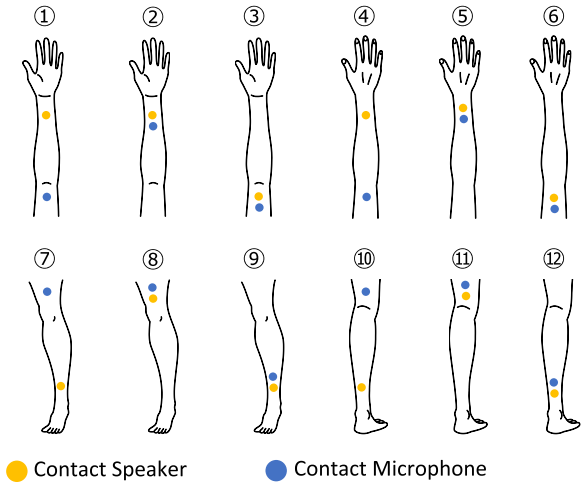

Fig. 8 Wearing position of device.

Table 1 Recognition result at each wearing position [%].

	Position	Precision	Recall
Arm	1	95.2	95.1
	2	99.8	99.8
	3	99.3	99.3
	4	98.2	98.2
	5	100.0	100.0
	6	97.9	97.7
Leg	7	71.5	71.5
	8	99.9	99.9
	9	99.7	99.7
	10	78.4	78.2
	11	96.2	96.0
	12	99.2	99.2

Table 2 Data acquired order of evaluation.

Order	Context	Situation
1	Sitting	Posture
2	Sitting Tightening Muscles	Posture
3	Typing	Behavior
4	Writing	Behavior
5	Scrolling	Mobile
6	Tapping	Mobile
7	Strong Tapping	Mobile
8	Wristwatch	Mobile
9	Wristwatch Tightening Muscles	Mobile
10	Drawer	Environment
11	Standing	Posture
12	Standing Tightening Muscles	Posture
13	Door	Environment
14	Twisting Tap	Environment
15	Refrigerator	Environment
16	Switch	Environment
17	Window	Environment
18	Going Up/Down Stairs	Behavior
19	Walking	Behavior
20	Jogging	Behavior
21	Bicycling	Behavior

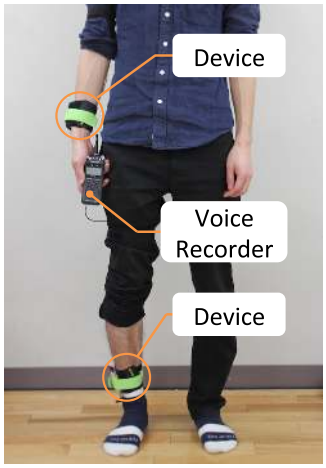

Fig. 9 Wearing position of devices.

Table 3 Result of training/testing from the same data set [%].

Subject	Day 1				Day 2			
	1st		2nd		1st		2nd	
	P[1]	R[2]	P	R	P	R	P	R
A	94.0	91.7	91.4	89.2	94.6	93.7	87.5	85.5
B	95.6	94.8	98.2	98.1	98.2	98.0	96.9	96.6
C	96.4	96.2	89.9	85.2	92.6	91.7	95.8	94.5
D	95.9	95.4	98.0	97.5	86.9	85.9	94.5	93.3
E	92.3	90.7	93.5	91.9	90.3	86.5	90.3	87.0
F	93.7	93.3	94.4	92.7	93.0	89.6	95.8	94.8
G	89.9	88.2	85.9	82.4	92.9	91.6	88.1	86.0
H	92.9	91.9	98.6	98.5	90.4	87.3	92.9	91.9
I	92.9	91.9	97.4	96.0	87.5	84.1	85.8	81.6
J	94.6	94.3	93.0	93.0	97.3	97.1	95.2	94.5
Average	93.8	92.8	94.0	92.5	92.4	90.6	92.3	90.6

[1]Precision [2]Recall

nition rate, we concluded that the contact microphone and the contact speaker should be worn close to each other.

5.2 Gesture Recognition

5.2.1 Procedure

We evaluated the proposed method. We assumed that our method would be integrated into an existing wearable device such as a smartwatch, and would be used in daily life for interface and life log. We assumed four situations in daily life: interaction with mobile device, interaction to environment, posture, and behavior. From these situations, we pick up contexts to be recognized in daily-life, as shown in **Table 2**. We add some contexts with tightening muscles since one of our targets is to detect tightening muscles to enhance the application of user interfaces and life-logging. For example, wristwatch tightening muscles is used to confirm the content of a notification on a smartwatch when both hands are busy and sitting/standing with tightening muscles is used as confidential communication and life log of stress. We used an iPhone 6s to acquire interaction with mobile device contexts. The door was a hinged door. The switch was a typical one to turn on/off the light. The window was a sliding window. The 10 subjects were 22 to 27-year-old males and females. They conducted 21 contexts for approximately one minute each. Table 2 shows the data acquired order of the contexts. We defined these 21 contexts as one set. The subjects did two sets in one day without detaching and re-attaching the device. Moreover, the subjects did two sets in another day. Although the wearing position of the device in the second day was approximately the same as that of the first day, we did not precisely specify the wearing position because we could confirm how the frequency spectrum was changed by the difference in the wearing position. We acquired 40 data sets in total (2 sets × 2 days × 10 subjects).

Figure 9 shows the wearing position of the devices. Based on the previous section result and wearability, we set the posterior wrist (position 5) and the posterior ankle (position 12) as the wearing position of the device. From the preliminary experiment, we confirmed that the signal from position 5 cannot be detected by the position 12 microphone, and the signal from position 12 cannot be detected by the position 5 microphone. Therefore, the subjects wore the device on the arm and leg at the same time. The voice recorder recorded in stereo. The left channel recorded the input from the leg microphone, and the right channel recorded the input from the arm microphone. We calculated features for the inputs from two microphones and combined the features for recognition.

From the preliminary experiment, we confirmed that the frequency characteristics slightly changed with time in the first few minutes after putting on the device. Thus, we did the experiments after the subjects had been wearing the device for 10 minutes. The recording sampling rate was 96 kHz, and the number of samples for FFT was 8,192 without any overlaps. Thus, the sampling rate for acquiring features was 11 Hz (96,000/8,192).

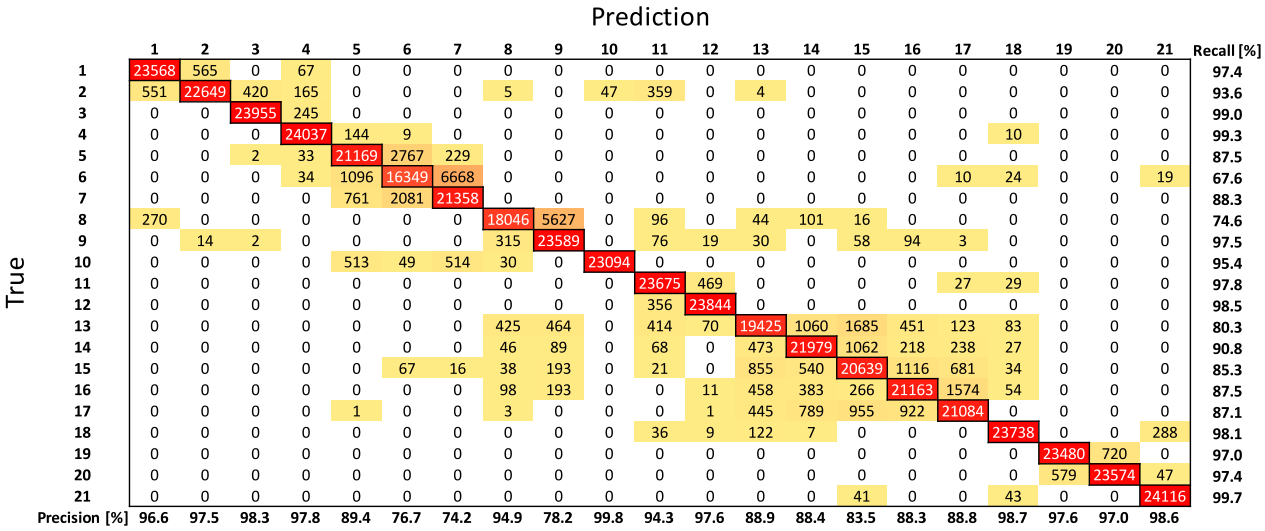

Fig. 10 Confusion matrix of training/test from the same data set.

Table 4 Recognition result of training/testing from other data set [%].

Subject	Day 1				Day 2				Train Day 1 Test Day 2		Train Day 2 Test Day 1	
	Train 1st Test 2nd		Train 2nd Test 1st		Train 1st Test 2nd		Train 2nd Test 1st					
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
A	63.4	65.0	58.6	56.3	85.0	82.3	72.0	70.8	32.4	34.4	23.6	21.8
B	29.2	37.5	63.1	65.5	68.3	70.2	71.7	69.6	32.5	35.7	34.7	32.8
C	27.8	29.1	29.1	28.2	60.3	59.7	76.8	71.8	15.0	15.3	9.2	12.4
D	51.5	54.6	55.6	58.4	50.8	39.5	50.7	53.2	7.3	14.3	10.4	14.7
E	42.8	41.1	35.2	41.4	25.4	24.4	22.0	20.7	23.3	16.0	15.5	13.2
F	43.9	46.8	72.6	69.2	58.5	63.6	65.9	63.7	12.3	11.9	15.2	16.7
G	48.2	47.1	60.8	59.6	39.7	34.7	35.0	36.9	13.0	14.1	4.4	5.4
H	66.0	51.6	68.0	48.7	37.9	49.0	71.6	63.2	24.0	22.0	16.4	14.4
I	15.7	25.9	22.3	24.8	31.6	33.7	33.7	40.1	10.0	15.9	16.8	10.6
J	74.0	69.8	48.6	57.6	64.8	62.5	66.0	62.2	18.0	23.0	25.3	19.6
Average	46.3	46.9	51.4	51.0	52.2	52.0	56.5	55.2	18.8	20.3	17.2	16.2

5.2.2 Result

We set 10% of the acquired data as training data and the rest as testing data. The average recognition rate is shown in **Table 3**. As shown in this table, the recognition rate was approximately 90% in all data sets. **Figure 10** shows the confusion matrix of all subjects and all data sets. The label numbers in Fig. 10 correspond to the numbers in Table 2. When the user keeps a static condition with/without tightening muscles (label 1, 2, 11, and 12), the precision and recall are 96.5% and 96.8%, respectively. When the user makes a gesture with/without tightening muscles (label 6–9), although the recognition rate is lower than the static condition, the precision and recall are 81.0% and 82.0%, respectively. With regard to the rest of the contexts, although there are some variations in the classification in interaction to environment contexts (label 13–17), the precision and recall are 93.5% and 92.6%, respectively. From the result, we consider that the proposed method can recognize not only the user's physical movements but also the user's internal states.

We also considered completely separating the training data set and testing data set. **Table 4** shows the result of training/testing by using another data set. As shown in this table, when we trained/tested by using the other data set in the same day, the recognition rate dropped to approximately 50% because the basic frequency spectrum was changed by the deviation in the position of the device caused by activities with large movements, such as jogging and bicycling. **Figure 11** shows frequency spectrums of sitting for all data sets of subject E, as an example. To grasp the approximate shape of the spectrum, we divided it into 20 blocks and calculated the mean at each block. This graph shows the means at 20 blocks. As shown in this figure, the basic frequency spectrum differed depending on the data set.

Figure 12 shows the confusion matrix of training/testing from the other data set in the same day. The recognition result is scattered in general. However, going up/down stairs, walking, jogging, and bicycling (label 18–21) were relatively accurately recognized because these contexts involved large movements, and the frequency change was large and characteristic compared to the other contexts.

When we trained and tested by using the other day's data set, the recognition rate was approximately 20%. The basic frequency spectrum differed from that of the other day because of the deviation in the wearing position and the pressure of wearing the device.

5.2.3 Revised Method

We considered a revised method to improve the recognition rate when we used the other data set for training and testing. We consider that the problem is that the basic frequency spectrum differs depending on the data set. Therefore, we set the first context (sitting) of each data set as the reference and focused on the change from the reference. Moreover, to detect minor changes in the spectrum, we divided each frequency spectrum bin by the corresponding reference bin. Then, we changed the magnifica-

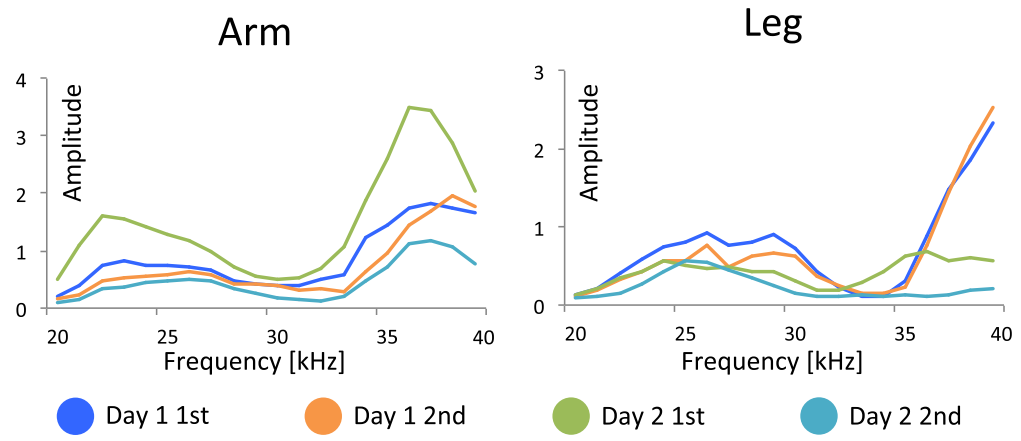

Fig. 11 Frequency spectrums of sitting for each data set (20–40 kHz).

Prediction

True \ Prediction	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Recall [%]
1	8494	808	2834	2433	2187	1926	0	81	32	3592	1165	614	0	340	728	225	672	702	0	0	47	31.6
2	758	10020	703	1351	508	257	625	74	525	2866	215	2201	1366	98	162	1324	224	214	0	36	3353	37.3
3	2043	877	10171	1002	1333	684	79	675	0	4026	487	0	235	1285	1858	134	1281	0	0	0	710	37.8
4	88	843	487	14389	2496	760	662	412	5	3446	13	0	750	583	0	290	626	139	0	672	219	53.5
5	1166	11	36	2288	7479	4386	3779	34	0	4025	166	438	607	375	515	0	768	631	0	0	176	27.8
6	1714	0	0	1205	5755	5903	5888	0	0	3112	672	2	484	520	465	141	973	46	0	0	0	22.0
7	1224	0	40	1429	5505	5459	7583	3	85	3011	672	1	701	83	362	0	722	0	0	0	0	28.2
8	728	0	597	242	118	151	81	10287	3327	612	242	1102	2153	898	2204	2587	964	525	0	0	62	38.3
9	109	0	610	162	0	0	62	2227	12382	89	363	1350	2968	944	1288	1831	1506	969	0	0	20	46.1
10	818	0	1723	644	654	504	972	0	69	19600	0	216	201	145	153	368	58	650	0	90	15	72.9
11	172	238	4	0	0	115	651	1229	1584	1470	9178	1794	3887	830	962	2396	261	1569	0	0	540	34.1
12	0	598	38	442	0	2	715	672	0	2202	1790	14296	1409	953	1133	229	514	1887	0	0	0	53.2
13	660	76	349	12	0	581	338	611	456	0	406	777	14081	1032	1270	1487	2220	2280	0	0	244	52.4
14	654	0	54	0	0	416	323	791	1	6	719	1317	2227	10701	3039	498	4454	1313	0	0	367	39.8
15	926	100	45	49	162	669	253	1102	459	373	247	769	1626	2624	10691	2017	4225	541	0	0	2	39.8
16	16	90	245	0	415	752	562	256	987	0	786	679	1825	384	2010	12028	5063	744	0	0	38	44.7
17	350	4	605	0	197	487	544	859	183	9	170	414	1747	1203	1159	753	17554	593	0	0	49	65.3
18	16	0	23	11	5	25	65	5	182	37	16	12	567	106	190	22	47	25076	0	0	475	93.3
19	0	0	0	0	0	0	0	0	0	0	0	0	0	7	38	0	1	30	23130	3674	0	86.0
20	0	23	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	642	1490	24713	6	91.9
21	0	2783	0	0	0	0	0	0	0	3	7	11	1	0	0	0	4	2541	0	0	21530	80.1
Precision [%]	42.6	60.8	54.8	56.1	27.9	25.6	32.7	53.3	61.1	40.4	53.0	55.0	38.2	46.3	37.9	45.7	41.7	61.0	93.9	84.7	77.3	

Fig. 12 Confusion matrix of training/testing from other data set.

Table 5 Recognition result by using revised method [%].

Subject	Day 1				Day 2				Train Day 1 Test Day 2		Train Day 2 Test Day 1	
	Train 1st Test 2nd		Train 2nd Test 1st		Train 1st Test 2nd		Train 2nd Test 1st					
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
A	72.7	69.1	59.1	51.1	79.0	75.0	70.5	65.8	28.7	17.6	36.0	37.2
B	39.3	43.8	66.2	63.2	66.3	68.6	68.7	59.4	25.1	27.8	38.1	28.3
C	39.9	43.6	29.8	33.2	80.2	73.5	63.3	64.5	22.5	17.5	17.4	15.0
D	56.9	58.4	64.5	66.9	54.7	46.4	33.0	31.6	10.7	14.3	9.4	16.6
E	39.0	40.7	36.7	42.5	23.8	27.4	22.9	21.6	13.5	12.1	15.1	16.0
F	80.2	81.4	69.9	68.3	63.6	68.2	61.2	59.5	10.4	10.8	13.0	19.0
G	58.2	45.5	57.1	52.6	29.3	28.7	30.3	31.0	13.9	7.5	15.8	11.9
H	70.6	74.7	58.1	54.4	45.5	52.1	39.8	44.7	12.0	11.3	34.2	21.7
I	12.8	15.7	15.1	19.4	31.4	26.3	30.2	32.0	13.4	15.5	14.3	17.8
J	73.2	67.6	45.9	46.4	69.7	67.6	65.9	63.0	16.3	14.7	23.6	23.8
Average	54.3	54.1	50.2	49.8	54.4	53.4	48.6	47.3	16.7	14.9	21.7	20.7

tion, and the minor change became major. **Table 5** shows the recognition results of the revised method.

Comparing Table 4 and 5, although there are improvements in some data sets and with some subjects, significant improvement is not observed by this method.

5.3 Discussion

When we used the other data set in the same day or the data set in the other day for training/testing, the recognition rate drastically decreased because the acquired frequency characteristic was different even when the posture was the same. We consider that there were three reasons for the change of the frequency characteristic: the deviation of device position, the change of device wearing strength, and the change of human-body characteristic. We plan to quantitatively clarify the influence for each factor and to develop a recognition algorithm to adopt these factors.

As practical ideas for tackling these problems, we fix the wearing position and the wearing strength of the device as far as possible whenever the user wears the device. Concretely, we plan to develop a smaller device for our method and integrate it into existing smartwatches. The wearing position of the device is fixed because he/she wears the device in almost the same position when-

ever the user wears the smartwatch. The wearing strength of the device is also fixed because the wearing strength is determined by selecting one of the holes at intervals placed on the watchband. The user selects the same hole and it means that the wearing strength and the position of device also become the same. If we can confirm the day-to-day change of the body characteristic in a future investigation, we resolve this problem by calibration when the user wears the device for the first time in the day. The system acquires his/her body characteristic of the day and utilizes the training data considering the body characteristic of the day.

Usually, the sensor sampling rates in conventional methods using accelerometers and EMG sensors are 20–100 Hz and 2 kHz, respectively. On the other hand, the sampling rate of our method is 96 kHz, whose calculation cost is more than that of the conventional method. Therefore, to solve this problem, we plan to mount a digital signal processor for data processing on the device.

Although our proposed method requires to adhere the device to the user's skin, we reduce the burden on the user by integrating the proposed device into the existing wearable device, such as a smartwatch. EMG sensors also need to be attached to the user's skin. However, it is difficult to recognize the physical movements by using them. Our proposed method is superior in terms of detecting not only physical movement but also the internal state at the same time by the sensor adhering to the skin.

We plan to develop the smaller device for out method and integrate it into the existing wearable device, such as a smartwatch. We also investigate the possibility of the implementation of the proposed method by using the built-in speaker and microphone of existing smartwatches.

Although we only focused on the ultrasonic range sound in this study, an audible-range sound, such as muscle activity sound, possibly is effective in some situations. We plan to combine the audible-range sound and ultrasound considering the situations.

The detected wave by the microphone consisted of the surface wave and the reflected wave. Separating these waves is useful for more precise recognition. We will develop our method and utilize the characteristics of these waves.

In this study, we used a limb for the wearing position of the device, because it is the most moving part in the human body. We plan to adapt the proposed method to other parts of the body, such as the face and abdominal part.

6. Conclusion

In this study, we proposed a gesture recognition method using active acoustic sensing. Our method transmits ultrasound to a user's body and recognizes his/her gestures by utilizing the detected ultrasound that propagates through the body. The proposed method can recognize the user's gestures by utilizing not only the physical movement but also the user's internal state, such as muscle activity. We evaluated 21 contexts for 10 subjects, and the evaluation results confirmed that when we set 10% of the data as training data and the rest as testing data, the precision and recall are 93.1% and 91.6%, respectively. When we used the other data set in the same day for training/testing, the precision and recall are 51.6% and 51.3%, respectively.

Acknowledgments This research was partly supported by a Grant in aid for Japan Society for the Promotion of Science Fellows Number 15J04608 and Precursory Research for Embryonic Science and Technology (PRESTO) from the Japan Science and Technology Agency.

References

[1] Audacity, available from ⟨http://www.audacityteam.org/⟩ (accessed 2017-01-22).
[2] Guidelines for the Safe Use of Ultrasound: Part Ⅱ - Industrial and Commercial Applications - Safety Code 24, available from ⟨http://www.hc-sc.gc.ca/ewh-semt/pubs/radiation/safety-code_24-securite/index-eng.php⟩ (accessed 2017-01-22).
[3] Hitohada Gel, available from ⟨http://www.exseal.co.jp/english/creative/hitohada.html⟩ (accessed 2017-01-22).
[4] Kinect for Windows, available from ⟨https://developer.microsoft.com/en-us/windows/kinect⟩ (accessed 2017-01-22).
[5] Leap Motion, available from ⟨https://www.leapmotion.com/?lang=en⟩ (accessed 2017-01-22).
[6] NDT Resource Center, Introduction of Ultrasonic Testing, available from ⟨https://www.nde-ed.org/EducationResources/CommunityCollege/Ultrasonics/cc_ut_index.htm⟩ (accessed 2017-01-22).
[7] Weka 3: Data Mining Software in Java, available from ⟨http://www.cs.waikato.ac.nz/ml/weka/⟩ (accessed 2017-01-22).
[8] Amma, C., Krings, T., Böer, J. and Schultz, T.: Advancing Muscle-Computer Interfaces with High-Density Electromyography, *Proc. SIGCHI Conference on Human Factors in Computing Systems* (*CHI 2015*), pp.929–938, ACM (2015).
[9] Brenner, M.C. and Fitzgibbon, J.J.: Surface Acoustic Wave Touch Panel System (1987). US Patent 4,644,100.
[10] Cook, J.L., Khan, K.M., Kiss, Z.S., Purdam, C.R. and Griffiths, L.: Prospective Imaging Study of Asymptomatic Patellar Tendinopathy in Elite Junior Basketball Players, *Journal of Ultrasound in Medicine*, Vol.19, No.7, pp.473–479 (2000).
[11] Fan, M. and Truong, K.N.: SoQr: Sonically Quantifying the Content Level inside Containers, *Proc. 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp 2015*), pp.3–14, ACM (2015).
[12] Fukui, R., Watanabe, M., Shimosaka, M. and Sato, T.: Hand Shape Classification in Various Pronation Angles Using a Wearable Wrist Contour Sensor, *Advanced Robotics*, Vol.29, No.1, pp.3–11 (2015).
[13] Gupta, S., Morris, D., Patel, S. and Tan, D.: Soundwave: Using the Doppler Effect to Sense Gestures, *Proc. SIGCHI Conference on Human Factors in Computing Systems* (*CHI 2012*), pp.1911–1914, ACM (2012).
[14] Laput, G., Brockmeyer, E., Hudson, S.E. and Harrison, C.: Acoustruments: Passive, Acoustically-Driven, Interactive Controls for Handheld Devices, *Proc. SIGCHI Conference on Human Factors in Computing Systems* (*CHI 2015*), pp.2161–2170, ACM (2015).
[15] Lewin, P.A. and Ziskin, M.C.: *Ultrasonic Exposimetry*, CRC Press (1992).
[16] Mokaya, F., Lucas, R., Noh, H.Y. and Zhang, P.: Myovibe: Vibration Based Wearable Muscle Activation Detection in High Mobility Exercises, *Proc. 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp 2015*), pp.27–38, ACM (2015).
[17] Mujibiya, A., Cao, X., Tan, D.S., Morris, D., Patel, S.N. and Rekimoto, J.: The Sound of Touch: On-body Touch and Gesture Sensing Based on Transdermal Ultrasound Propagation, *Proc. 2013 ACM International Conference on Interactive Tabletops and Surfaces* (*ITS 2013*), pp.189–198, ACM (2013).
[18] Murao, K., Terada, T., Yano, A. and Matsukura, R.: Evaluating Sensor Placement and Gesture Selection for Mobile Devices, *Information and Media Technologies*, Vol.8, No.4, pp.1154–1165 (2013).
[19] Ono, M., Shizuki, B. and Tanaka, J.: Touch & Activate: Adding Interactivity to Existing Objects Using Active Acoustic Sensing, *Proc. 26th Annual ACM Symposium on User Interface Software and Technology* (*UIST 2013*), pp.31–40, ACM (2013).
[20] Pirkl, G., Stockinger, K., Kunze, K. and Lukowicz, P.: Adapting Magnetic Resonant Coupling Based Relative Positioning Technology for Wearable Activitiy Recogniton, *Proc. 2008 International Symposium on Wearable Computers* (*ISWC 2008*), pp.47–54, IEEE (2008).
[21] Rahman, T., Adams, A.T., Zhang, M., Cherry, E., Zhou, B., Peng, H. and Choudhury, T.: Bodybeat: A Mobile System for Sensing Non-Speech Body Sounds, *Proc. 12th Annual International Conference on Mobile Systems, Applications, and Services* (*MobiSys 2014*), pp.2–13, ACM (2014).
[22] Ren, Z., Yuan, J. and Zhang, Z.: Robust Hand Gesture Recognition Based on Finger-Earth Mover's Distance with a Commodity Depth

Camera, *Proc. 19th ACM International Conference on Multimedia*, pp.1093–1096, ACM (2011).
[23] Sato, M., Poupyrev, I. and Harrison, C.: Touché: Enhancing Touch Interaction on Humans, Screens, Liquids, and Everyday Objects, *Proc. SIGCHI Conference on Human Factors in Computing Systems* (*CHI 2012*), pp.483–492, ACM (2012).
[24] Stiefmeier, T., Ogris, G., Junker, H., Lukowicz, P. and Troster, G.: Combining Motion Sensors and Ultrasonic Hands Tracking for Continuous Activity Recognition in a Maintenance Scenario, *Proc. 2006 International Symposium on Wearable Computers* (*ISWC 2006*), pp.97–104, IEEE (2006).
[25] Takemura, K., Ito, A., Takamatsu, J. and Ogasawara, T.: Active Bone-Conducted Sound Sensing for Wearable Interfaces, *Proc. 24th Annual ACM Symposium on User Interface Software and Technology* (*UIST 2011*), pp.53–54, ACM (2011).
[26] Watanabe, H., Terada, T. and Tsukamoto, M.: Ultrasound-based Movement Sensing, Gesture-, and Context-recognition, *Proc. 2013 International Symposium on Wearable Computers* (*ISWC 2013*), pp.57–64, ACM (2013).
[27] Yatani, K. and Truong, K.N.: BodyScope: A Wearable Acoustic Sensor for Activity Recognition, *Proc. 2012 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp 2012*), pp.341–350, ACM (2012).

Editor's Recommendation

The authors propose a gesture recognition method using active acoustic sensing that transmits acoustic signals for recognizing the user's state by analyzing the response. Novelty and originality of the proposed method are sufficiently high. We expect that the result of this paper inspires further researches in not only gesture recognition but also activity recognition, and recommend this paper to the Journal of Information Processing (JIP).

(Chairman of SIGUBI Kazushige Ouchi)

Hiroki Watanabe is in the doctoral course at Graduate School of Engineering, Kobe University, Japan. He received B.Eng. and M.Eng. degrees from Kobe University in 2012, 2014, respectively. Mr. Watanabe is working on wearable computing and ubiquitous computing. He is a member of IPSJ.

Tsutomu Terada is an Associate Professor at Graduate School of Engineering, Kobe University, Japan. He received B.Eng. M.Eng. and Ph.D. degrees from Osaka University in 1997, 1999, and 2003, respectively. Professor Terada is working on wearable computing, ubiquitous computing, and entertainment computing. He is a member of IEEE, IPSJ, and IEICE.

Masahiko Tsukamoto is a Professor at Graduate School of Engineering, Kobe University, Japan. He received B.Eng. M.Eng. and Ph.D. degrees from Kyoto University in 1987, 1989, and 1994, respectively. Professor Tsukamoto is working on wearable computing and ubiquitous computing. He is a member of ACM, IEEE, and IPSJ.