A Method for Embedding Context to Sound-based Life Log

Watanabe, Hiroki
Terada, Tsutomu
Tsukamoto, Masahiko

(Citation)
Journal of Information Processing,22(4):651-659

(Issue Date)
2014-10

(Resource Type)
journal article

(Version)
Version of Record

Journal of Information Processing Vol.22 No.4 651–659 (Oct. 2014)

[DOI: 10.2197/ipsjjip.22.651]

Regular Paper

A Method for Embedding Context to Sound-based Life Log

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

Received: November 20, 2013, Accepted: May 17, 2014

Abstract: Wearable computing technologies are attracting a great deal of attention on context-aware systems. They recognize user context by using wearable sensors. Though conventional context-aware systems use accelerometers or microphones, the former requires wearing many sensors and a storage such as PC for data storing, and the latter cannot recognize complex user motions. In this paper, we propose 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. The system recognizes gestures on the basis of the volume of the received sound and the Doppler effect. The former indicates the distance between the neck and wrists, and the latter indicates the speed of motions. We combine the gesture recognition by using ultrasound and conventional MFCC-based environmental-context recognition to recognize complex contexts from the recorded sound. Thus, our approach substitutes the wired or wireless communication typically required in body area motion sensing networks by ultrasounds. Our system also recognizes the place where the user is in and the people who are near the user by ID signals generated from speakers placed in rooms and on people. The strength of the approach is that, for offline recognition, a simple audio recorder can be used for the receiver. Contexts are embedded in the recorded sound all together, and this recorded sound creates a sound-based life log with context information. We evaluate the approach on nine gestures/activities with 10 users. Evaluation results confirmed that when there was no environmental sound generated from other people, the recognition rate was 86.6% on average. When there was environmental sound generated from other people, we compare an approach that selects used feature values depending on a situation against standard approach, which uses feature value of ultrasound and environmental sound. Results for the proposed approach are 64.3%, for the standard approach are 57.3%.

Keywords: wearable computing, gesture recognition, environment recognition, ultrasonic sound, life log, location recognition, person recognition

1. Introduction

In wearable computing, the human activity recognition technologies enable novel activity-aware services (Lifelog, Gesture operation, etc.). Typical sensors used for activity recognition include wearable motion sensors sensing limb movements (e.g., accelerometers, inertial measurement units) [1], [2], or on-body microphones sensing sound generated by the user's activities [3]. Wearable motion sensors may sense the user's own activities precisely, but cannot be used to recognize the activities of surrounding users. Furthermore, to recognize complex activities, data from motion sensors placed on each part of the body should be integrated. On-body networking poses its own challenges, such as energy use in wireless networks, or clothing integration wired networks. A wearable microphone, on the other hand, captures the sound generated by human activities performed by the user itself, or other persons in his/her neighborhood. Some activities that are challenging to recognize from a network of motion sensors may be easier to identify from the sound if it is itself very characteristic. However, this approach cannot be used to recognize precise body movements.

In this paper, we propose a context recognition method that uses only sound to sense both gestures and environment sounds. In our system, the user wears a receiver comprising a microphone on chest and wears one or more small speakers to generate ultrasonic sound on the wrists. The system recognizes gestures on the basis of the volume of the received sound that determines the distance between the speaker and the microphone. Additionally, our method uses the Doppler effect to estimate the speed of motions. Our system also recognizes the place where the user is in and the people who are near the user by ID signals generated from speakers placed in rooms and on people. The strengths and the novelty of the proposed approach are that a single microphone can be used for gesture recognition, for the recognition of places, and for the sensing of neighboring users. Contexts are therefore embedded to the sound-based life-log data. In other words, user contexts are recorded in only one audio file. The technical novelty is that a user wear small ultrasound speakers on both wrists, and we use a volume of ultrasound and shift of ultrasound frequency caused by Doppler effect for gesture recognition. Since the receiver (microphone) and transmitters (speakers) are within a body of one person, the system does not need any wired or wire-

[1] Graduate School of Engineering, Kobe University, Kobe, Hyogo 657–8501, Japan
[2] 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

less communication function. Finally, for offline context recognition, the receiver can be a simple off-the-shelf audio recorder.

We implemented a prototype of our proposed method using an audio recorder for offline context recognition. We evaluated the method on nine activities with 10 users. When there was environmental sound generated from other people, the recognition rate decreases 30% on average. We propose a feature selection method to increase the recognition rate in such condition. We also evaluated the recognition rate of person recognition, location recognition, and activity recognition in a daily scenario. We discuss how this method could be used for online recognition as well.

This paper is organized as follows. In Section 2, we describe related work. In Section 3, the recognition method is presented. The implementation is described in Section 4. The recognition rate of our method is discussed in Section 5. We discuss possible problems in Section 6. Finally, Section 7 concludes our research.

2. Related Work

2.1 Acceleration Based Recognition

Many studies on gesture recognition with accelerometers have been reported. Bao et al. developed and evaluated a new method to detect physical activities from data acquired from five small biaxial accelerometers worn simultaneously on different parts of the body [1]. Murao et al. evaluated recognition accuracy for 27 kinds of gestures with nine accelerometers and nine gyroscopes on a board and demonstrated the differences in recognition accuracy by changing the number, positions, and kinds of sensors and the number and kinds of gestures [4]. In these studies, it is difficult to recognize the contexts related to environmental sound such as talking because a user's surroundings cannot be recognized with only accelerometers. Moreover, each accelerometer should have a wireless/wired communication function to communicate with the devices for data integration. There are some studies using mobile phones to recognize context [5], [6]. Although these researches achieved practical context recognition in daily life since many people have already carried with their own mobilephones, they cannot recognize detailed context including user arm motions unless combined with additional acceleration sensors.

2.2 Other Sensor Based Recognition

There are many context recognition methods using various types of sensors. Some activity recognition methods using acoustic features are studied [7], [8], [9], [10]. In these studies, they use only environmental sound recognition.

Pirkl et al. demonstrate motion tracking systems by using magnetic field [2]. Starner et al. present a wearable device to control home automation systems via hand gestures [11]. The gestures are recognized using a small camera worn like a pendant. Mattmann et al. present a garment using strain sensors to recognize upper body postures [12]. Strain sensors are attached to the back region of a tight-fitting clothing. These sensors measure strain in the garment caused by different body movements and allow distinguishing and allow identifying/recognizing between a predefined set of body postures. Naya et al. studied nurse's routine activity recognition system [13]. Nurses have to memorize what they did in a day to communicate with each other and should not make mistakes such as giving an unnecessary dose of medicine. This system recognizes nurses' activities with an accelerometer and their locations with RF-ID receivers. Ward et al. described nine consecutive contexts in woodwork (sawing, hammering, filing, drilling, grinding, sanding, opening a drawer, tightening a vice, and turning a screwdriver) that are recognized by using microphones and three-axis accelerometers mounted at two positions on the user's arms [3]. However, in these studies, when multiple sensors are placed on the user's body separately, there is a need to have wireless/wired communication functions to communicate with the devices for storing/processing data.

2.3 Ultrasonic Based Recognition

Ultrasonic transmissions are used for tracking people [14], [15]. User's location is recognized by Doppler shift or times-of-flight of sound pulses from an ultrasonic transmitter to receivers placed at known positions. Combining ultrasonic with other sensors, hand tracking systems are proposed [16], and gesture recognition methods by using ultrasonic Doppler shift are developed [17], [18]. In these studies, ultrasonic is used for only location tracking or gesture recognition in fixed place.

3. Recognition Method

Our method requires the user to wear small ultrasonic speakers to recognize the user's motions, placed on that person's wrist. A speakers emitting the person's ID are worn on the chest. Additional speakers for location recognition are set-up on office desks. In our evaluation we use a simple voice recorder, worn on the chest, as receiver and analyze the data offline. **Figures 1** and **2** show the device configurations for gesture recognition, location recognition, and person recognition. To recognize the difference between the right hand and the left hand gestures, the frequency for both hands is set to be different. More specifically, in this research, we use 19,000 Hz for the frequency of the left wrist and 20,000 Hz that of the right wrist. In this paper, we evaluate only one user wearing speakers at the same time for gesture recognition. When some users wear the speakers for gesture recognition,

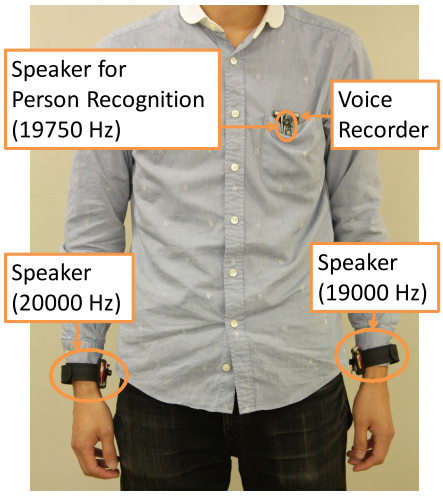


Fig. 1 Device configuration.

we slightly changed the usage frequency. To recognize the difference between location and person, we use 19,500 Hz for the frequency of the location recognition and 19,750 Hz for that of the person. Each location/person has a unique ID.

The voice recorder records life-log data including conversation with other people, the ultrasonic sound from both hands, that of the location, and that of the ID of surrounding people at the same time. In our system, the receiver (voice recorder) and the transmitters (small speakers) are within the body of one person. Since the speakers just transmit ultrasonic, and the voice recorder just records the sound, they do not need to communicate with other devices for data fusing.

We use ultrasonic sound for the speakers to recognize gestures because it is noisy when we use audible range sound. In addition, environmental sound and the sound from the speaker can be easily separated to recognize the environment and gestures at the same time. The user can select the target of recognition by selecting worn devices, as shown in **Table 1**. The sampling frequency of sound is set to 44.1 kHz. The flow of recognition is shown in **Fig. 3**.

3.1 Pre-processing of Sound Data

The features of the signal of the sound are that the low frequency is a large amplitude spectrum and the high frequency is a small amplitude one. To correct the bias in this frequency, the high frequency is emphasized on the basis of the following formula [19]. N is the number of samples (N = 4,096). α_n is n-th sound data ($n = 1, \cdots, N$) before emphasizing the high frequency. α'_n is the n-th sound data after emphasizing the high frequency.

$$\alpha'_n = \alpha_n - 0.97\alpha_{n-1} \tag{1}$$

The beginning and the end of the waveform cut out for processing is discontinuous and inconvenient when using Fourier transform. Therefore, we multiply a window function to smooth the boundary. In this paper, we use a Hamming window [20]. β_n is the n-th data ($n = 1, \cdots, N$) of the Hamming window. γ_n is the n-th data after multiplying by sound data and window function, as shown in the following formula.

$$\beta_n = 0.54 - 0.46 \cos \frac{2\pi n}{N-1} \tag{2}$$

$$\gamma_n = \alpha'_n \times \beta_n \tag{3}$$

To separate environmental and ultrasonic sound, the frequency spectrum is calculated by using a fast Fourier transform (FFT). The vertical axis shows the power of sound, and the horizontal axis shows the frequency. We can obtain the power of each frequency.

3.2 Activity Recognition

Generally, a feature value is calculated from sensor data to recognize contexts effectively. In this paper, we use the Mel-frequency cepstral coefficient (MFCC) as the feature value for environmental sound. MFCC is a feature value that emphasizes the important frequencies of the human auditory system. MFCC has 20 dimensions, and 12 low-level element dimensions are generally taken as the feature value.

As feature values for motion recognition, we use the mean and the variance of the volume of ultrasonic sound $V(T)$. The mean $\mu(T)$ and the variance $\sigma^2(T)$ of 10 samples in the past are calculated on the basis of the following formula [21]. T means the current time.

$$\mu(T) = \frac{1}{10} \sum_{t=T-9}^{T} V(t) \tag{4}$$

$$\sigma^2(T) = \frac{1}{10} \sum_{t=T-9}^{T} \{V(t) - \mu(T)\}^2 \tag{5}$$

The feature value is four dimensions, which are the mean and the

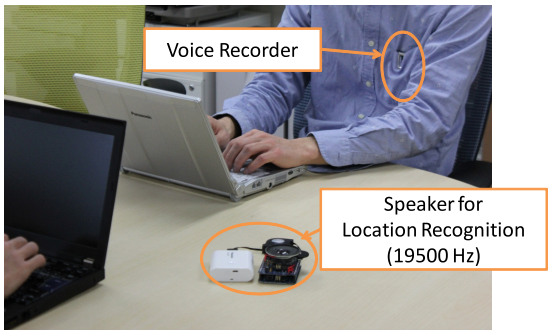


Fig. 2 A speaker for location recognition.

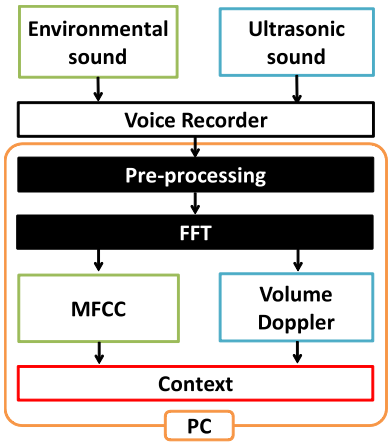


Fig. 3 Flow of recognition process.

Table 1 Selecting devices.

Worn device	Environment	Gesture	Location	Person
Recorder	○			
Recorder + Wrists	○	○		
Recorder + Location	○		○	
Recorder + Person	○			○
Recorder + Wrists + Location	○	○	○	
Recorder + Wrists + Person	○	○		○
Recorder + Wrists + Location + Person	○	○	○	○

variance for the volume with 19,000 Hz and 20,000 Hz. Note that the volume of ultrasonic sound includes a certain margin in a frequency. This is because the Doppler effect changes the frequency of generated sound. Concretely, in this research, we use 15 Hz as the margin for calculating feature values.

To consider the Doppler effect as feature values in order to recognize user motions, we use the frequency that has a maximum volume within 50 Hz before and after the frequencies 19,000 Hz and 20,000 Hz. We can obtain the sequence of the frequency peaks, and then we use the mean and variance of 10 samples preceding the frequency peak as four dimensional feature values.

The above-mentioned feature values are 20 dimensions in total $X(T) = (x_1(T), \cdots, x_{20}(T))$. Since these feature values are at different scales, they are normalized on the basis of the following formula [21]. $Z(T)$ means normalized data. M is the mean. S is the standard division. T is time.

$$Z(T) = \frac{X(T) - M}{S} \tag{6}$$

For the recognition, we employ the euclidean distance between training data $Z_i = (z_{i1}, \cdots, z_{ij}, \cdots, z_{i20})$ and the unknown data $Z = (z_1, \cdots, z_j, \cdots, z_{20})$ calculated by using the following formula, which is commonly used in activity recognition.

$$d_i = \sqrt{\sum_{j=1}^{20} (z_{ij} - z_j)^2} \tag{7}$$

The label of data that has minimum distance becomes recognition result (K-nearest neighbor algorithm, K = 1). Though more complicated methods such as SVM (Support Vector Machine) or decision trees can be used for more accurate recognition, we use this simple method to know the difference between the cases with/without our method.

3.3 Location and Person Recognition

An ultrasonic ID is sent from speakers set up in an environment and on people. The frequency used for the speakers is 19,500 Hz to recognize a location and 19,750 Hz to recognize a person, while other frequencies can be used if we would like to use many places and people. We employ the amplitude modulation for the modulation scheme of ultrasonic ID for locations and persons. Each speakers has a unique ID which is composed of 1 and 0. Speakers transmit ultrasound while each bit of ID is 1, and do not emit ultrasound while the bit of ID is 0. When the system analyzes a captured data, if the volume of the sound of corresponding frequency is larger than the threshold, the system recognizes that the received data is 1. An ID consists of a header part (ex. 1010) and main ID part. The system reads the main ID portion only after recognizing the header. In this experiment, we use 4 bits for main ID, and each location and person has unique main ID such as 0001 for Location A and 0010 for Person A. We can freely expand the number of possible main IDs by adding more bits to main ID part. The system can recognize the location or person by comparing obtained IDs with IDs assigned to each location and person. In this research, we set an interval of 1 pulse every 0.5 seconds.

Fig. 4 Appearence of devices.

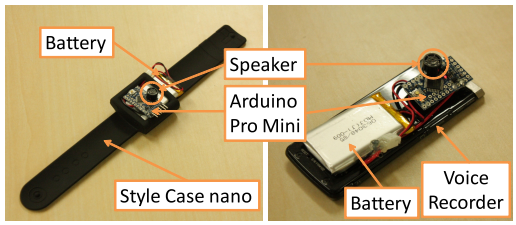


Fig. 5 Appearence of small devices.

4. Implementation

We implemented a prototype based on the proposed method. The device, worn on the wrists, consisted of a speaker, an Arduino Nano ver. 3.0, a SparkFun Electronics lithium polymer battery (3.7 V, 2,000 mAh), and a SparkFun Electronics LiPo Charger/Booster, as shown in top of **Fig. 4**. These parts were sewn to a wristband in order to be attached/detached easily. The Arduino Nano generates a square wave via a speaker. The device used for location and person recognition consisted of a speaker, an Arduino Uno R3, Wave Shield ver. 1.1, and a Panasonic lithium ion battery QE-QL201X-W (5 V, 5,400 mAh), as shown at the bottom of Fig. 4. The wave file of ultrasonic ID was made with a PC by using Audacity, which is a free software for editing and recording audio, and then it was stored on an SD card. It was played by using Arduino and Wave shield. We also implemeted a small prototype as shown in **Fig. 5**. The small type device consisted of a small speaker, an Arduino Pro Mini and a lithium polymer battery. We attach this unit to Style Case nano, which is a wristband for iPod nano, to wear on wrist. Because Style Case nano winds around the wrist easily, we can attach/detach a device easier. A speaker unit for person recognition is attached to the back side of a voice recorder. In this paper, we used a former prototype. The voice recorder used was Sony ICD-TX50. The recording mode was LPCM (44.1 kHz, 16-bit). We used Microsoft Visual C# to develop the software that analyzed the acquired data. The PC used was a Panasonic Let's Note CF-S9LYYKDS (CPU: Core i5, Memory: 4 GB).

5. Evaluation

5.1 Without Sound Emitted by Other People

We first assume an ideal environment where there is no sound emitted by other users since our method uses a recognition method with auditory analysis. Note that most evaluations for

sound recognition in conventional researches have used such ideal environments. We evaluated nine contexts: sitting, walking, running, eating, brushing teeth, cleaning, washing hands, typing, and talking, which are assumed to be daily indoor activities. These contexts are selected from Activities of Daily Living and Instrumental Activities of Daily Living [22], which are difficult to with just one type of sensor. Cleaning is assumed to be done with a vacuum cleaner, brushing teeth is assumed to be done not with an electric toothbrush but with an ordinary one, and eating is assumed to be eating spaghetthi with a fork. We evaluated the context recognition accuracy in changing the selection of the feature values. The 10 subjects were 21 to 23-year-old males and females. The sampling rate was 10 Hz for both environmental and gesture recognition. The subjects did the action for each context for approximately 45 seconds, and we used 50 samples of each context performed by the subjects themselves for the training data. Three hundred other samples that were not used for the training data for each context were used for evaluation data. The appearance of the experiment is shown in **Fig. 6**. The vacuum cleaner used in the experiment was a Dyson DC26. A Dell SK-8115 was used as the keyboard. The K-nearest neighbor algorithm was used for recognition (K = 5), and the window size was 4,096 for environmental recognition and 10 for gesture recognition.

Table 2 shows the result of the average recognition rate for all 10 subjects. The accuracy of recognition was 74.4% on average with MFCC, 78.7% with Volume, 48.7% with the Doppler effect. When using only MFCC, the recognition rate of the contexts without regular sounds such as running, eating, and talking was not so high. In comparison, the recognition result of feature sounds such as sitting (silence) and cleaning (loud) was good. Brushing teeth was recognized at a high recognition rate when using Volume as a feature value. This is because when the user brushed his/her teeth, only their dominant hand was very near to the recorder. When using only Doppler as a feature value, it was difficult to recognize these contexts. By combining two kinds of feature values, there was a 13.4% increase in the average accuracy compared with using just one kind of feature value. Each feature value mutually complemented the low accuracy of the other feature. Combining three kinds of feature values, the recognition rate was 86.6% on average. As seen from the above, combining different kinds of feature values is effective. When there is no other environmental sound, we should use all three kinds of feature values.

Fig. 6 Snapshots on experiment without sound emitted by others.

Compared with the system developed by Bao et al. [1], they use five accelerometers and recognition accuracy is 89.71% in walking, 94.78% in sitting, 87.68% in running, 85.27% in brushing teeth, 97.49% in working on computer, 88.67% in eating or drinking, 96.41% in vacuuming. The average recognition rate of these seven contexts are 91.43%. This result is slightly higher than proposed method, however they need to wear five accelerometers on each part of the body. Moreover, this method cannot recognize the sound related contexts such as talking. The training method is similar to the conventional approaches. However, the most important difference between conventional methods and our method is hardware configulation and restriction on communication.

5.2 With Sound Emitted by Others

We evaluated our method when there was environmental sound from other people to investigate how this sound influences the recognition rate. We selected five environmental sounds (typing, washing, brushing, cleaning, and talking), which are considered to different types of environmental sounds. The subjects performed each action in the presence of environmental sound from others, as shown in **Table 3**, for approximately 45 seconds. Because it is not easy to consider situations with these environmental sounds when the user is running, we did not evaluate with running. The 10 subjects were 21 to 23-year-old males and females, and the sampling rate was 10 Hz. The training data was the same data used in the previous experiment. The appearance of the experiment is shown in **Fig. 7**.

First, we used the same recognition method used in the previous experiment. As the result shown in **Table 4**, the recognition rate decreased on the whole. The average recognition rate of all contexts in the presence of environmental sound from others was 57.3%. This is because such sound negatively affected the recognition results. In particular, when there were cleaning sounds, the recognition rate greatly fall. This is because the sound of the vacuum cleaner was very loud compared with the other sounds. Therefore, other sounds were erased by the vacuum cleaner, and the recognition rate by using MFCC with included feature values greatly worsened.

To recognize correctly in the presence environmental sound from others, we propose an improved recognition method. Note

Table 2 Recognition rate without sound emitted by others [%].

Combination of feature values	Sitting	Walking	Running	Eating	Typing	Brushing	Washing	Cleaning	Talking	Average
MFCC	90.6	80.4	60.6	63.4	70.1	66.6	77.6	100	60.5	74.4
Volume	85.1	77.3	68.4	64.0	81.4	95.5	79.9	73.8	82.8	78.7
Doppler	97.5	60.1	62.7	41.2	24.8	41.2	39.1	48.0	23.7	48.7
MFCC + Volume	93.2	90.5	68.2	77.3	86.1	78.3	86.6	100	70.7	83.4
Volume + Doppler	88.7	67.8	69.8	70.2	91.6	84.1	80.0	69.6	89.2	79.0
MFCC + Doppler	97.7	76.6	73.6	70.2	82.6	67.3	82.3	100	66.1	79.6
All	98.9	82.8	76.0	82.0	91.8	78.5	88.0	100	81.5	86.6

Table 3 Combination of user behaviors and sounds emitted by others.

User's action	Environmental sound of others
Sitting	Typing, Washing hands, Brushing teeth, Cleaning, Talking
Walking	Typing, Washing hands, Brushing teeth, Cleaning, Talking
Running	-
Eating	Typing, Talking
Typing	Typing, Washing hands, Brushing teeth, Cleaning, Talking
Brushing teeth	Typing, Washing hands, Brushing teeth, Cleaning, Talking
Washing hands	Typing, Brushing teeth, Cleaning, Talking
Cleaning	Typing, Washing hands, Brushing teeth, Talking
Talking	Typing, Washing hands, Brushing teeth, Cleaning, Talking

Table 4 Average accuracy with sound emitted by others [%].

Other's environmental sound	Combinations of feature values	Sitting	Walking	Eating	Typing	Brushing	Washing	Cleaning	Talking	Average
Typing	MFCC	17.8	65.3	17.1	72.3	29.8	59.8	100	54.5	52.1
	Volume	37.7	53.6	32.6	55.9	75.2	58.8	50.7	60.5	53.1
	Doppler	89.6	52.2	45.6	23.8	29.7	31.8	43.4	13.2	41.2
	MFCC + Volume	16.9	80.7	32.3	67.0	48.9	76.7	100	62.9	60.7
	Volume + Doppler	59.4	59.5	40.9	66.5	67.2	59.2	53.3	77.7	60.5
	MFCC + Doppler	47.0	68.6	27.4	91.5	38.1	64.0	100	57.4	61.7
	All features	50.9	74.3	37.8	83.7	50.7	82.1	99.9	71.9	68.9
Washing	MFCC	17.6	57.3	-	43.6	17.6	-	99.3	50.5	47.7
	Volume	43.1	55.1	-	45.1	77.4	-	36.7	61.8	53.2
	Doppler	78.3	57.7	-	24.0	20.8	-	44.0	16.6	40.3
	MFCC + Volume	23.5	75.0	-	44.6	39.6	-	99.5	62.8	57.5
	Volume + Doppler	57.2	63.4	-	55.8	63.7	-	48.5	82.8	61.9
	MFCC + Doppler	54.4	68.5	-	66.6	19.6	-	99.3	54.9	60.5
	All features	71.4	73.9	-	68.5	41.8	-	99.2	73.4	71.3
Brushing	MFCC	43.3	71.5	-	58.1	34.9	68.8	100	54.7	61.6
	Volume	58.9	60.2	-	44.0	67.7	74.5	38.1	68.1	58.8
	Doppler	95.7	59.1	-	28.7	33.7	40.0	45.0	10.2	44.6
	MFCC + Volume	58.5	85.0	-	53.1	51.9	84.8	100	65.4	71.2
	Volume + Doppler	70.4	66.8	-	62.6	59.8	75.2	51.2	79.5	66.5
	MFCC + Doppler	74.2	73.1	-	83.4	41.2	74.6	99.9	55.2	71.6
	All features	82.5	77.9	-	78.9	51.6	90.7	99.8	73.5	79.3
Cleaning	MFCC	0.0	0.0	-	0.0	0.0	0.9	-	2.6	0.6
	Volume	41.8	67.0	-	47.8	76.4	63.2	-	59.4	59.3
	Doppler	65.1	48.1	-	24.4	40.4	34.6	-	5.3	36.3
	MFCC + Volume	0.0	0.1	-	1.6	0.1	1.7	-	8.4	2.0
	Volume + Doppler	39.6	56.2	-	59.5	73.4	63.0	-	29.2	53.5
	MFCC + Doppler	0.0	6.2	-	1.4	0.0	4.4	-	2.2	2.4
	All features	0.6	16.6	-	11.2	1.9	5.6	-	7.2	7.2
Talking	MFCC	16.9	32.6	17.8	17.2	11.8	46.4	99.9	64.6	38.4
	Volume	59.0	67.8	37.9	57.5	76.7	80.9	41.4	60.5	60.2
	Doppler	88.2	55.9	42.0	26.1	26.5	39.1	50.7	11.0	42.4
	MFCC + Volume	23.2	46.2	30.2	20.3	29.0	72.5	99.9	76.5	49.7
	Volume + Doppler	59.7	62.9	45.2	66.0	66.3	76.9	55.0	84.5	64.6
	MFCC + Doppler	28.1	56.5	30.8	37.4	17.1	54.9	99.9	68.1	49.1
	All features	34.6	64.0	37.6	45.1	34.5	79.4	99.7	84.3	59.9

Fig. 7 A snapshot of experiment with sound emitted by others.

that the purpose of improved method is not raise the average recognition rate but prevent radical fall of recognition rate in particular situations.

The basic idea of improving the method is to select the feature values depending on the situation. When there is no sound emitted by others, we should use all three features because of the previous evaluation result that gives the best recognition rate. When there is sound emitted by others, we should use two feature values without MFCC, which indicates only user's motion. This is because MFCC contains others' sound that negatively affects the recognition while user's motion (volume and doppler) is not affected by others' sound. The flow of improved recognition

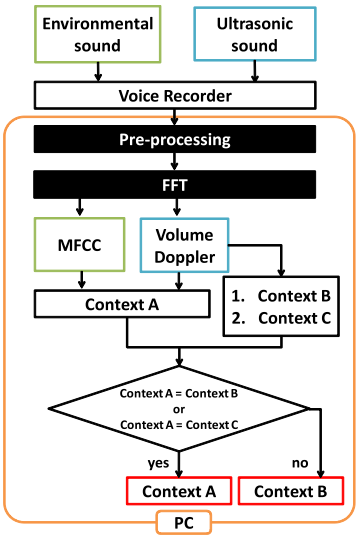


Fig. 8 Recognition process considering sound emitted by others.

Table 5 Accuracy in our revised method.

Sound emitted from others	Accuracy [%]	
	Conventional method	Revised method
Typing	68.9	64.7
Washing	71.3	66.1
Brushing	79.3	71.7
Cleaning	7.2	51.0
Talking	59.9	67.8
Average	57.3	64.3

method is shown in **Fig. 8**. The method calculates one recognition result (context A) that uses three feature values, and calculates two recognition candidates (context B and C) which use two feature values (volume of ultrasound and Doppler) using k-NN algorithm (k = 5). A first order of recognition result is set to context B, and a second order of recognition result is set to context C. When context A is the same as context B or context C, we consider that context A is not affected by others' sound. Therefore, we use context A as the final result. Otherwise, we consider that context A is affected by the sound emitted by others. Thus the method uses context B as the final result.

Applying this mechanism, although the recognition accuracies of several methods slightly decreased, the recognition rates improved when there was cleaning sound and talking. In particular, when others' sound is Cleaning, there is significant improvement of recognition rate that is approximately 40% of recognition accuracy. Moreover, the recognition rate of all contexts was 64.3%, an improvement of 7.0%, as shown in **Table 5**. The purpose of improved method is not raise the average recognition rate but prevent greatly fall of recognition rate in particular situation. Therefore, from the above discussion, we consider that the improved method was effective.

5.3 Evaluation of Location and Person Recognition

We evaluated the accuracy of location and person recognition. The speaker that generates ultrasonic ID was set up on desks each of five rooms. The subject was one person, who sats on a chair in front of the desk for 30 seconds and then moved to another room in specific order. Repeating this for two sets, we confirmed the accuracy of the location recognition. Though all 10 places were able to be recognized accurately, one place had been recognized as a different location in the middle of recognition. The reason is considered to be that the ultrasonic ID did not get recognized accurately due to influences that got in the way of the ID acquisition.

Table 6 Action scenario.

# of context	Context	Location	Environmental sound from others
1	Typing	Room A	Talking
2	Eating	Room B	Typing
3	Sitting		
4	Talking with A	Room A	
5	Cleaning		
6	Talking with B	Room B	
7	Typing		Typing
8	Brushing	Room C	
9	Washing		

When we used five speakers in five locations in the same room at the same time, each ID interfered with each other and it was not able to recognize IDs accurately. Therefore, in the case where we use multiple speakers in the distance at which sounds interfered with each other, it is necessary to shift the sending interval of an ID so as not to overlap with that of others. In location and person recognition, the error of recognition does not become a serious problem since there are many chances for recognition in a conversation or a stay of the room.

We also evaluated person recognition. Three subjects wore a speaker that sent ultrasonic ID on their chest, and one of them wore voice recorder too on his/her chest. The three subjects spoke face to face for approximately two minutes. All three subjects were able to be recognized in approximately one minute. In this system, the more the number of subjects increases, the more time is needed to recognize all subjects. This is because the pulse of the ID sending is set to 0.5 seconds, which is long. By shortening the pulse, the time taken until recognition is able to be shortened.

Compared with the system developed by Ward et al. [14], 95% of raw readings lie within 14 cm of the true position. They can recognize where the user is in the room. However, in this system, the receivers have to be set up at known positions. In our system, though the system cannnot recognize where the user was in the room precisely, it can recognize in which room the user was, and it is easy to set up speakers for location recognition.

5.4 Action Scenario

To confirm the practicality of the proposed method in daily life, we evaluated daily actions along scenarios, as shown in **Table 6**. The movement between contexts was entirely walking. The subjects were three males. The voice recording was done for approximately 10 minutes. The user wore a voice recorder on his chest and ultrasonic speakers on his wrists. Two subjects, person A and B, wore the speaker that generates an ultrasonic ID for person recognition on their chest. The speakers that generate an ultrasonic ID for location recognition were set up in three places, as shown in **Fig. 9**. **Figure 10** shows the ground truth of actual action.

Figure 11 shows the recognition result. Recognized results are normalized by majority decision for 10 seconds. As this fig-

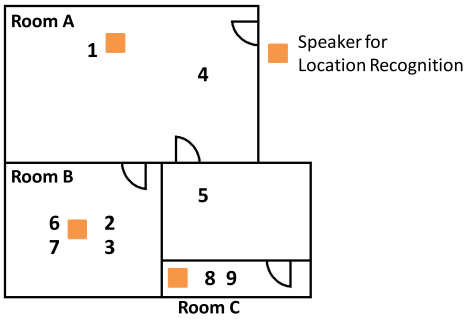


Fig. 9 Location of speakers and actions.

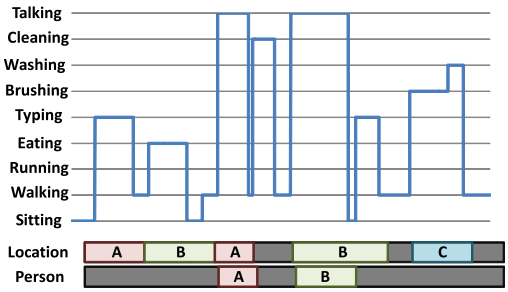


Fig. 10 Ground truth of actual action.

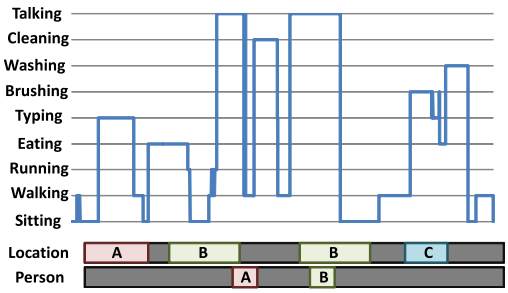


Fig. 11 Scenario result.

ure shows, context was mostly recognized correctly. However, typing in Room B was not recognized correctly. The reason is considered to be that the keyboard in Room B was different from the one used when taking training data. Typing in Room A was correctly recognized because the keyboard in Room A was the same one used when taking training data. Four places out of the five places were able to be recognized correctly. When the user talked with person A, location information was not recognized. This is because Room A was larger than Room B and Room C, as shown in Fig. 9. When the user was in the Room A, the volume is not enough. All activities of the user in Room B were done near the speaker, however the user talked with person A at a point a little far from the speaker in Room A. Therefore, the volume of ultrasonic sound was not enough, and could not be used for recognition. To recognize more correctly, we should enlarge the output from the speaker or set up speakers in the room within the range where the sound does not overlap.

Persons A and B were both able to be recognized correctly. When the user was using the vacuum cleaner, misidentification of the location and person occasionally occurred. This is because the sound of a vacuum cleaner contains various frequencies. 19,500 Hz, used for the location recognition, and 19,750 Hz, used for the person recognition, are also contained in vacuum cleaner sound. We can solve this problem by setting the appropriate threshold or setting a higher frequency that is not included in cleaner sounds for the speakers.

6. Discussion

Conventional sound recording systems have a privacy problem, which records the sound we do not want to be heard by other people such as the sound when the user was in a rest room. Our system tags the recorded sound and can remove/encrypt the part of the recorded sound based on the tags. We discuss how this method could be used for online recognition as well. When use devices having microphone and data processing capacity, such as smartphone, instead of simple voice recorder, we consider that the system can recognize online.

In this study, as shown in Section 3.2, we set 50 Hz as a margin for Doppler effect. In other words, we need 100 Hz per one wrist. Thus, we need 200 Hz for one person. Since we use 44,100 Hz as a sampling rate, the maximum frequency that can be detected when we use Fourier transform is 22,050 Hz. In this paper, since we use the frequency of 19,000 Hz or more as a ultrasound, the range of frequency which can be detected is 3,050 Hz between 19,000 Hz to 22,050 Hz. Considering the above, our system can recognize 15 persons at the same time by a simple calculation. However, since the volume of wrist speakers is not so loud, we consider that it does not become a significant restriction. This is because the sound from a people standing nearby the user is small enough not to be detected.

When there are two or more users using the same frequency, we consider to change the frequency for wrist speakers automatically. Concrete algorithm will be proposed in future, but there is a simple algorithm that wrist speakers stop transmitting ultrasound at regular time intervals, and the system confirms whether there is a sound of the same frequency in neighborhood. When there is a sound of the same frequency, wrist speakers change the frequencies to be used.

Since most humans cannot hear sound at around 20,000 Hz [23], the sound used in our system is almost imperceptible to humans. On the other hand, many of the non-human animal, such as dog or cat, have wider audible range compared with human. The frequency used in this paper may be uncomfortable for pets from the fact that there are many products to keep vermin away by using ultrasonic. If we use the system with animals, we should carefully consider the setting of volume and frequency of ultrasonic.

Ultrasonic sound is reflected by the wall, thus if the users use this system in narrow space, reflection can negatively affect to the recognition result. In this evaluation, experiment was done in enough large room, therefore there seems no effects of reflection.

In this study, we assume the user wear the speakers on wrists and voice recorder on chest, since we consider the hands the part of the body most often moved, and chest is the most stable position in the body.

7. Conclusion

In this paper, we propose an ultrasound-based context recognition method. Our method utilizes small speakers that output ultrasonic for adding contexts to the recorded sound of a voice

recorder. When there is no sound emitted by others, the system could recognize the context correctly at an average of 86.6%. When there was sound emitted by others, the average was 57.3%, while that was 64.3% with the revised method. Moreover, we evaluate location, person, and daily activity recognition, and confirmed the effectiveness of the proposed method.

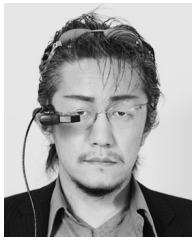
Acknowledgments This research was supported in part by a Grant in aid for Precursory Research for Embryonic Science and Technology (PRESTO) from the Japan Science and Technology Agency.

References

[1] Bao, L. and Intille, S.S.: Activity recognition from user-annotated acceleration data, *Proc. 2nd International Conference on Pervasive Computing* (*PERVASIVE '04*), Vol.3001, pp.1–17 (2004).
[2] Pirkl, G., Stockinger, K., Kunze, K. and Lukowicz, P.: Adapting magnetic resonant coupling based relative positioning technology for wearable activitiy recogniton, *Proc. 12th International Symposium on Wearable Computers* (*ISWC '08*), pp.47–54 (2008).
[3] Ward, J.A., Lukowicz, P., Troster, G. and Starner, T.E.: Activity recognition of assembly tasks using body-worn microphones and accelerometers, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.28, No.10, pp.1553–1567 (2006).
[4] Murao, K., Terada, T., Yano, A. and Matsukura, R.: Evaluating gesture recognition by multiple-sensor-containing mobile devices, *Proc. 15th International Symposium on Wearable Computers* (*ISWC '11*), pp.55–58 (2011).
[5] Iso, T. and Yamazaki, K.: Gait analyzer based on a cell phone with a single three-axis accelerometer, *Proc. 8th International Conference on Human-Computer Interaction with Mobile Devices and Services* (*MobileHCI '06*), pp.141–144 (2006).
[6] Duong, T.V., Bui, H.H., Phung, D.Q. and Venkatesh, S.: Activity recognition and abnormality detection with the switching hidden semi-markov model, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR 2005*), Vol.1, pp.838–845 (2005).
[7] Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J. and Sorsa, T.: Computational auditory scene recognition, *Proc. 27th International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP '02*), Vol.2, pp.1941–1941 (2002).
[8] Eronen, A.J., Peltonen, V.T., Tuomi, J.T., Klapuri, A.P., Fagerlund, S., Sorsa, T., Lorho, G. and Huopaniemi, J.: Audio-based context recognition, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.14, No.1, pp.321–329 (2006).
[9] Stager, M., Lukowicz, P. and Troster, G.: Implementation and evaluation of a low-power sound-based user activity recognition system, *Proc. 8th International Symposium on Wearable Computers* (*ISWC '04*), Vol.1, pp.138–141 (2004).
[10] Al Masum Shaikh, M., Molla, M.K.I. and Hirose, K.: Automatic life-logging: A novel approach to sense real-world activities by environmental sound cues and common sense, *Proc. 11th International Conference on Computer and Information Technology* (*ICCIT '08*), pp.294–299 (2008).
[11] Starner, T., Auxier, J., Ashbrook, D. and Gandy, M.: The gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring, *Proc. 4th International Symposium on Wearable Computers* (*ISWC '00*), pp.87–94 (2000).
[12] Mattmann, C., Amft, O., Harms, H., Troster, G. and Clemens, F.: Recognizing upper body postures using textile strain sensors, *Proc. 11th International Symposium on Wearable Computers* (*ISWC '07*), pp.29–36 (2007).
[13] Naya, F., Ohmura, R., Takayanagi, F., Noma, H. and Kogure, K.: Workers' routine activity recognition using body movements and location information, *Proc. 10th International Symposium on Wearable Computers* (*ISWC '06*), pp.105–108 (2006).
[14] Ward, A., Jones, A. and Hopper, A.: A new location technique for the active office, *IEEE Personal Communications*, Vol.4, No.5, pp.42–47 (1997).
[15] Muller, H.L., McCarthy, M. and Randell, C.: Particle filters for position sensing with asynchronous ultrasonic beacons, *Location-and Context-Awareness*, pp.1–13 (2006).
[16] Ogris, G., Stiefmeier, T., Junker, H., Lukowicz, P. and Troster, G.: Using ultrasonic hand tracking to augment motion analysis based recognition of manipulative gestures, *Proc. 9th International Symposium on Wearable Computers* (*ISWC '05*), pp.152–159 (2005).
[17] Kalgaonkar, K. and Raj, B.: One-handed gesture recognition using ultrasonic Doppler sonar, *Proc. 34th International Conference on Acoustics, Speech and Signal Processing* (*ICASSP '09*), pp.1889–1892 (2009).
[18] Gupta, S., Morris, D., Patel, S. and Tan, D.: SoundWave: Using the doppler effect to sense gestures, *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems* (*CHI '12*), pp.1911–1914 (2012).
[19] Loweimi, E., Ahadi, S.M., Drugman, T. and Loveymi, S.: On the Importance of Pre-emphasis and Window Shape in Phase-Based Speech Recognition, *Proc. 6th International Conference on Non-Linear Speech Processing* (*NOLISP '13*), pp.160–167 (2013).
[20] Rajput, S.S. and Bhadauria, D.S.: Comparison of Band-stop FIR Filter using Modified Hamming Window and Other Window functions and Its Application in Filtering a Mutitone Signal, *International Journal of Advanced Research in Computer Engineering & Technology* (*IJARCET*), Vol.1 (2012).
[21] Murao, K., Terada, T., Takegawa, Y. and Nishio, S.: A context-aware system that changes sensor combinations considering energy consumption, *Proc. 6th International Conference on Pervasive Computing* (*PERVASIVE '08*), pp.197–212 (2008).
[22] Activities of Daily Living: What Are ADLs and IADLs?, available from ⟨http://www.caring.com/articles/activities-of-daily-living-what-are-adls-and-iadls⟩.
[23] Frequency Range of Human Hearing, available from ⟨http://hypertextbook.com/facts/2003/ChrisDAmbrose.shtml⟩.

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

Tsutomu Terada is an Associate Professor at Graduate School of Engineering, Kobe University. He received B.Eng., M.Eng., and Ph.D. degrees from Osaka University in 1997, 1999, and 2003, respectively. He has been an Assistant Professor at Cybermedia Center of Osaka University and a Lecture in 2000 and 2005, respectively. He is currently investigating the wearable computing, ubiquitous computing, and entertainment computing. He is a member of IEEE and IEICE.



Masahiko Tsukamoto is a Professor at Graduate School of Engineering, Kobe University. He received B.Eng., M.Eng., and Ph.D. degrees from Kyoto University in 1987, 1989, and 1994, respectively. From 1989 to 1995, he was a research engineer of Sharp Corporation. From 1995 to 1996, he has been an Assistant Professor at the Department of Information Systems Engineering, Osaka University and since 1996, he has been an Associate Professor at the same department. He is currently investigating the wearable computing and ubiquitous computing. He is a member of eight learned societies, including ACM and IEEE.