Detecting Door Events Using a Smartphone via Active Sound Sensing

THILINA DISSANAYAKE, TAKUYA MAEKAWA*, DAICHI AMAGATA, and TAKAHIRO HARA, Osaka University, Japan

Event detection of indoor objects, including doors, has a wide variety of applications, including intruder detection, HVAC control, and surveillance of independently living elderly people. Hence, this has been the focus of multiple research projects in the UbiComp research community. Herein, we propose a method to accurately detect door events in an indoor environment, without the installation and maintenance costs of using distributed ubiquitous sensors. In particular, we recognize the events of multiple doors existing in the environment via active sound probing using a disused smartphone installed in the environment. We perform event recognition by fusing the analysis of the Doppler shift caused by the moving doors with the acoustic characteristics describing the open/close states of the doors acquired via impulse response. To accurately distinguish between the events of different doors via sound probing, our method employs the time-series analysis of the Doppler shift as well as the active sound probing using directional high-frequency sine waves and stereo sound recording. In addition, by incorporating prior knowledge about the state transitions of a door object into a recognition model, we attempt to improve the accuracy of event recognition. Moreover, our method is capable of recognizing walking activities of a person related to door events in the environment, which are necessary information for applications such as HVAC control that require information about both door events and human presence.

CCS Concepts: • **Human-centered computing** → **Ubiquitous computing**; • **Mathematics of computing** → *Kalman filters and hidden Markov models*; • **Computing methodologies** → *Supervised learning by classification*;

Additional Key Words and Phrases: Indoor context recognition, open/close event, active sound sensing, pattern recognition

ACM Reference Format:
Thilina Dissanayake, Takuya Maekawa, Daichi Amagata, and Takahiro Hara. 2018. Detecting Door Events Using a Smartphone via Active Sound Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 160 (December 2018), 26 pages. https://doi.org/10.1145/3287038

1 INTRODUCTION

Indoor context recognition is one of the most important research topics in the UbiComp research community because it is the basic technology behind various applications, such as lifelogging and context-aware applications. In particular, the techniques used to recognize the movement of indoor objects, such as when doors or windows in a room are opened or closed have a variety of applications, including intruder detection, HVAC control, and surveillance of independently living elderly persons. To date, many of the UbiComp studies have employed distributed sensors to observe such indoor events. In many cases, these studies employed sensors such as accelerometers, magnetometers, and state-change sensors (e.g., reed switches and magnets) to achieve indoor event recognition [21, 27, 30].

*This is the corresponding author

Authors' address: Thilina Dissanayake; Takuya Maekawa, takuya.maekawa@acm.org,maekawa@ist.osaka-u.ac.jp; Daichi Amagata; Takahiro Hara, Osaka University, Graduate School of Information Science and Technology, Suita, Osaka, 5650871, Japan.

2474-9567/2018/12-ART160 $15.00
https://doi.org/10.1145/3287038

However, because this approach requires attaching a separate sensor node to each object, their installation and maintenance costs are high. For example, frequent replacement of the node batteries or faulty sensor nodes places a large burden on users. When multiple nodes are installed in a house, the possibility of nodes becoming faulty is high, thereby indicating the need for frequent maintenance personnel visits to either fix or replace these nodes. Based on a public database of IoT-enabled houses, Kodeswaran et al. [10] revealed that, on an average, houses equipped with 14 - 100 sensor nodes require a maintenance personnel visit every 18 days. In addition, attaching sensor nodes to indoor objects negatively affect the aesthetic value of the artifacts [1]. Furthermore, a system based on the distributed door sensors is unable to detect whether a person using a door is entering or exiting the environment, making it inapplicable to applications related to door event detection such as HVAC control and surveillance that require information about both door events and human presence. This problem can be solved by adding an infrared-based motion sensor to the environment; however, the sensor network is further burdened by the increase in the initial price of the system while the system is prone to breakdowns. Furthermore, the sensing range of indoor environments can be quite complex as the motion sensor must deal with a considerable number of obstacles.

Therefore, a reliable method for indoor event recognition that works well with a small number of devices is required. Ohara et al. [19] proposed a method for door-based event detection using Wi-Fi channel state information (CSI). However, this method requires a specially modified Wi-Fi driver-installed PC with a special Wi-Fi module to be configured in the environment of interest. In addition, since the method relies on deep learning because manual feature design of CSI data is difficult, it requires huge amounts of training data containing more than 100 instances of open/close events for each door. Mahler et al. [18] proposed a smartphone-based home security system based on indoor event recognition. Currently, smartphones have become commodity devices and new models are released every year. Therefore, disused smartphones have considerably increased and are often left unattended, which makes it possible to employ these smartphones for indoor event recognition. In the method proposed by Mahler et al., a smartphone is either mounted on the door itself or to a wall near the latching mechanism of the door. When the smartphone is mounted on the door, the magnetic sensor data from the magnetometer is used to detect the open/close events of the door. Alternatively, when the smartphone is mounted to the wall, three-dimensional acceleration data from the accelerometer are used to detect the vibrations caused by the open/close events of the door. However, this method makes it difficult to increase the sensing range as it requires multiple smartphones to be mounted on/next to each door in an environment that has more than one door.

Herein, we also focus on an unattended disused smartphone to be used as a part of a door event recognition system. Since the average upgrade cycle of a smartphone in the U.S. is only 32 months[1], we can assume that a family of four buys a new smartphone every 8 months on average, indicating that a disused smartphone, which contains a variety of sensors, can be reused as a part of intelligent context-aware systems. In contrast to Mahler et al.'s method, we employ active sound sensing, which allows us to increase the sensing range of event recognition. Therefore, we propose installing a disused smartphone in a room to detect the events of the doors existing in the room. Indoor objects, including doors, have fixed states, e.g., the opened/closed states, and events, e.g., the open/close events, where the state transitions occur. Herein, the smartphone emits inaudible high-frequency sine waves (18 kHz and 20 kHz) from the speaker to avoid disturbing users and records the reflected wave using the inbuilt microphones. This enables us to capture the Doppler shift caused by the transition events of the door. In addition, a sweep signal is emitted periodically and from its impulse response, the acoustic characteristics of the environment comprising information related to the states of the objects in the surroundings are captured [26]. The acoustic characteristics of the environment differ according to the opened/closed state of the doors

[1]https://www.npd.com/wps/portal/npd/us/news/press-releases/2018/the-average-upgrade-cycle-of-a-smartphone-in-the-u-s–is-32-months—according-to-npd-connected-intelligence/

in the environment. By employing an impulse response based on a sine sweep, we attempt to recognize the states of the doors. Specifically, we repeatedly emit sine waves and sine sweeps from the smartphone (after each 10-s sine wave signal, the smartphone emits 0.05-s sine sweep signal) to obtain information related to both the Doppler shift and impulse response. In addition, the door event recognition method is robust against daily life noises such as voices because we focus only on a narrow high-frequency band to extract features. Besides door events, our method recognizes walking events of a person in relation to the door events using the Doppler shift, which is essential information for HVAC control applications.

In order to distinguish between the events of the different door objects, the proposed method uses the following information:

- The frequency and amplitude of the reflected sound containing information of the Doppler shift changes over time, which differs with the relative location of the object with respect to the smartphone. (This is comprehensively explained in Section 3.) Therefore, to model an event, we employ hidden Markov models (HMMs) [31], a widely used approach for modeling time-series data.
- Directionality of the sound waves emitted by a speaker changes with the frequency [24]. Therefore, by emitting a composite sine wave comprising low- and high-frequency components, we obtain information related to the relative location of the object with respect to the smartphone.
- Smartphones are commonly equipped with multiple microphones that also enable to identify the direction of a sound source.

We analyze the above acoustic information to recognize events/states of doors. Doppler shift can be observed over time, as mentioned above, so the events/states of each object can be modeled using HMMs. A left-to-right HMM is employed to model each event/state of each door. Moreover, we know that an "open" event only occurs when the door is first in the "closed" state, followed by the transition to the "opened" state. Similarly, a "close" event can only occur when the door is in the "opened" state, followed by the transition to the "closed" state. Based on the aforementioned understanding, we establish a grammar that is employed to decode input signals by the HMMs using the Viterbi algorithm [22].

Note that the information about the events/states of different doors contained in the recorded audio signals can possibly blend with each other. Thus, there is a risk that the recognition accuracy could deteriorate when the signal captured by the microphone (i.e., the feature vector sequences extracted from the signal) is directly fed into the model designed to predict the events/states of each object (a set of HMMs; HMM set). Here, we separate the input signal containing information regarding multiple doors into information about the events of each door before the signals are fed into the HMMs prepared for each object. To achieve this, we employ a discriminative classifier that is designed to recognize all door events in a given environment in order to transform the time-series signals obtained from the microphones into the time-series of the probabilities of the occurrence of each event. Thus, by feeding these time-series probabilities into the HMMs of each object, we can recognize the states/events of these objects. Furthermore, by incorporating the generative model, i.e., HMMs, we attempt to cope with sporadic noises included in the recorded sound, as well as capture temporal regularity of the events/states.

The contributions of this research are as follows:

- We propose a novel approach to recognize the states/events of indoor objects via active sound sensing. In addition, we propose a two-tier structure to recognize the events/states of multiple door objects in an environment. First, we detect the events/states using a discriminative classifier to separate the input signals of the events/states of multiple doors into information concerning the events of each door. This information is then fed into an HMM set prepared for each door, which decodes the input information using a grammar that defines the state transitions of the door. This is the first study that uses a smartphone to detect the events/states of the door objects via active sound sensing.

Table 1. Features of door event recognition methods

	installation cost	maintenance cost	coverage	privacy issue	dark	train	power	simultaneous	person	multiple persons
distributed	high	high	narrow	no	work	no	battery	yes	no	no
barometer	–	low	wide	no	work	yes	AC/battery	no	no	no
camera	–	low	wide	yes	not work	yes	AC	yes	yes	yes
RF	–	low	wide	no	work	yes	AC	no	yes	no
sound	reuse	low	wide	yes	work	yes	AC	no	yes	no

- To recognize the events/states of multiple objects with high precision, the proposed method is designed to employ the following: (1) the time-series analysis of the Doppler shift caused by the open/close events of the indoor objects, (2) the reflected waves from a composite signal comprising two sine waves with different frequencies, (3) the acoustic signals obtained from multiple microphones, and (4) the acoustic characteristics of the environment, which relate to the states of the indoor objects, obtained from the impulse response.
- This method can recognize walking events related to door events, which is necessary information to implement indoor applications such as HVAC control and surveillance.
- The data collected from different environments with different conditions is used to confirm the effectiveness and validity of the proposed method.

The remainder of this study is organized as follows: First, we introduce the studies related to context recognition for indoor objects. We then perform a preliminary investigation of active sound probing using smartphones. We design a method for recognizing door events via active sound sensing. Finally, we evaluate the proposed method using the sensor data obtained from real environments.

2 RELATED WORK

2.1 Monitoring Indoor Events Using Distributed Sensors

Event detection methods for indoor objects employ a large number of distributed sensors, such as accelerometers, RFID tags, switch sensors, and vibration sensors [16, 17, 21, 27, 30]. These sensors are attached to indoor objects to detect their events. Though high accuracy and fine-grained measurement of the events is guaranteed in the distributed sensor approach, the maintenance cost (e.g., battery and faulty node replacement) is usually high. Several researchers have attempted to resolve this issue by employing energy-harvesting sensors for monitoring buildings. Campbell et al. [3] proposed a method using a piezo-film to develop a vibration detector that generates an electrical current when vibrated and transmits a packet to report that vibration event. This sensor can be used to detect open/close events of a door, window, cabinet, or refrigerator door.

Table 1 summarizes the features of the approach based on the distributed door sensors. The price (related to installation cost) of current commercial devices is expensive (approximately 100 USD per device) because the current system mainly focuses on entrance doors for intrusion detection. The current system is inapplicable to other applications because it is unable to detect a person related to the door events. As mentioned in the introduction section, the maintenance cost is high when a number of nodes are installed.

2.2 Monitoring Indoor Events Using a Small Number of Sensors

Open/close events of doors have also been detected using barometric pressure sensors. Patel et al. [20] propose a method by employing pressure sensor units attached to HVAC air filter to detect pressure variations caused by open/close events of doors and room-to-room transitions. Wu et al. [32] employ a barometer in a smartphone

to observe a sharp change in the indoor pressure when a door event occurs in an environment with an HVAC system. However, this approach works well only in an environment with an HVAC system.

Shi et al. [25] recognize indoor situations such as "empty room", "opened room", and "walking person" by employing FM-radio signal receivers based on the fact that the propagation of radio waves is affected by changes in the environment. Wi-Fi channel information can also be used to detect the events and states of indoor everyday objects, such as doors and windows. Ohara et al. [19] propose a method to use a commodity Wi-Fi access point and a computer with a special Wi-Fi module inside a room to detect the changes that occur in Wi-Fi signal propagation during indoor events. In contrast, we attempt to recognize indoor events using a commodity smartphone.

Table 1 summarizes the features of the approaches based on the small number of sensors in addition to our approach (sound row). These approaches require training data to recognize events of multiple doors. We implement a smartphone application that supports data collection to efficiently collect labeled training data. For example, the application asks a user to use doors, e.g., reading out "Please open door A," and then the user uses the door according to the instructions.

The barometer-, camera-, RF-, and sound-based approaches require an AC power supply. To reduce installation and maintenance costs related to power supply, we can supply electricity via the easily available ceiling outlet or bulb socket in any room [12, 14].

The drawback of the barometer-based approach is that this approach is unable to detect a person related to the door events. In contrast, because the RF- and sound-based approaches can detect the person but are unable to distinguish between multiple persons, these methods can be applied to a variety of applications, such as HVAC control and surveillance of independently living elderly people. These approaches also cannot detect door events when another person is moving in the environment at the same time.

While the camera-based approach can detect multiple persons, this approach has critical issues related to privacy and dark environments. Although the barometer-, sound-, and RF-based approaches cannot detect door events that occur at the same time, the probability with which multiple doors are opened/closed at the same time is considerably low. While the sound-based approach also suffers from the privacy issue, our method works using only sound features of the inaudible high-frequency band. In order to further reduce the problems regarding the privacy, we suggest to extract the relevant features from the audio data before sending the data to a server where the recognition process runs.

As detailed above, the RF- and sound-based approaches do not have critical drawbacks and can be applied to a variety of applications. However, as mentioned in the introduction, the RF-based approach requires a specially modified Wi-Fi driver-installed PC with a special Wi-Fi module.

2.3 Context Recognition Using Active Sound Sensing

Many researchers make use of active sound sensing for context recognition. Gupta et al. [9] propose a method of recognizing hand gestures using the Doppler shift where they employ a tone in the range of 18 kHz as a pilot tone. Fu et al. [8] propose a method of using a 20 kHz sound wave to track exercises using the Doppler shift caused by the movements of the body.

Several mobile computing studies have employed sound beaconing to estimate relative positions to other devices [6, 33]. Active sound sensing has also been used to locate a smartphone user. Rossi et al. [23] propose a method to locate a smartphone based on active sound fingerprinting. The authors measure impulse response at each indoor reference point to train a classifier that estimates a user's current coordinates using observed impulse response. Tung et al. [29] also make use of active sound probing for indoor location tagging. Tachikawa et al. [26] employ active sound sensing as well as passive sound sensing by smartphones to estimate location semantics such as toilet and restaurant.

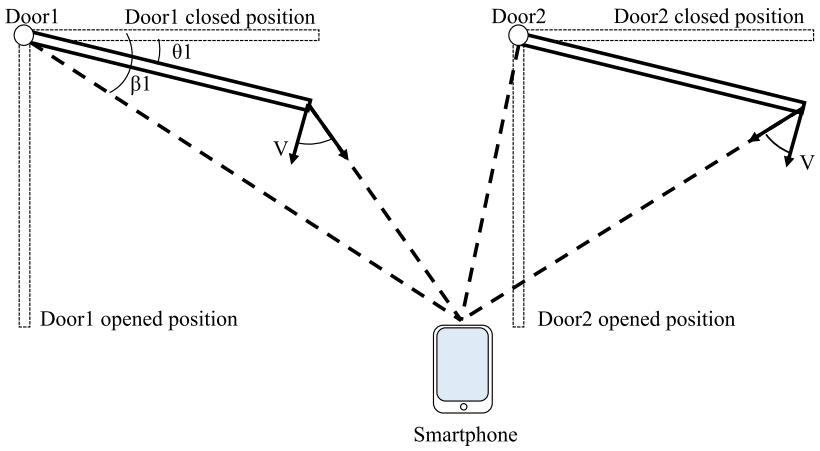

Fig. 1. Relation between the relative location of a moving door and its velocity component towards the smartphone

Similar to the above approaches, we also employ a sound wave with a constant frequency to detect the events of indoor objects. Note that our method has several features, which are mentioned in the introduction section including making use of composite signals and a grammar that defines state transitions of an object, to achieve accurate door event/state recognition.

2.4 Context Recognition Using Passive Sound Sensing

Passive sound sensing has also been used for context recognition. Tarzia et al. [28] used passive sound fingerprinting to locate a smartphone user by extracting sound fingerprints based on acoustic background spectrum of rooms. Maekawa et al. [13, 15] employed a wrist-worn device containing a camera and microphone to recognize object-based activities. Clarkson et al. [5] used a camera and a microphone attached to a chest strap to detect location related events such as entering an office, kitchen, or courtyard. Korpela et al. [11] employed a smartphone microphone to evaluate toothbrushing activities by analyzing sound events of toothbrushing.

3 INVESTIGATION OF ACTIVE SOUND SENSING

We investigated the characteristics of signals obtained via active sound sensing, and based on this investigation, we designed an indoor event recognition method in the next section.

3.1 Theory of the Operation of Event Detection Using Doppler Shift

To detect the open/close events of indoor objects, such as doors, we employed a well-known and well-understood phenomenon known as the "Doppler effect" or the "Doppler shift." This effect is defined as the shift of the observed frequency of a wave when the observer is moving relative to the wave source. In this research, the smartphone becomes both the observer and the wave source and the Doppler shift is caused by the moving doors in the environment. Imagine a scenario wherein a smartphone installed inside a room emits a wave with a constant frequency while recording the reflected wave at the same time. When a door event occurs in the environment, the resulting reflected wave then has a different frequency than the original frequency that the smartphone emitted. This frequency shift can be observed as a distortion in the FFT spectrum of the recorded sound. This distorted spectrum can then be analyzed to detect the door event and differentiate between the door events of multiple doors. In other words, when the door rotates around its hinge and moves toward the smartphone, it

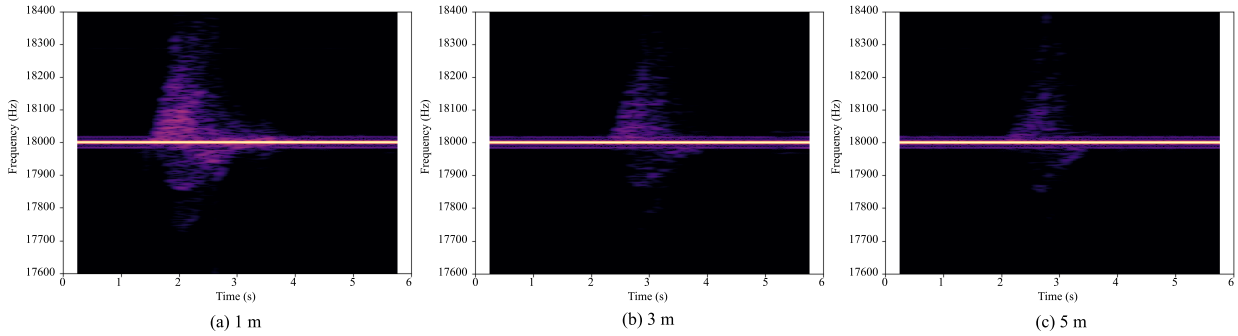

Fig. 2. FFT power spectrograms for a door-opening event when the smartphone is placed (a) 1 m from the door, (b) 3 m from the door, and (c) 5 m from the door

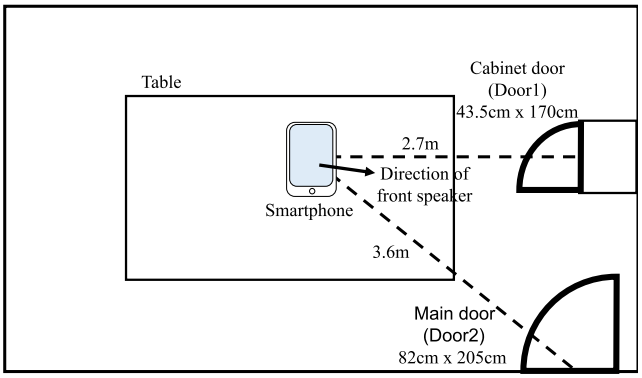

Fig. 3. Environment of preliminary experiment (1). The smartphone emits a sinusoidal sound wave with a frequency of 20 kHz. The directions of the front speaker and microphone are identical.

causes an increment in the recorded frequency, creating a positive shift in the FFT spectrogram. Similarly, when the door is moving away from the smartphone, it causes a decrement in the recorded frequency, creating a negative shift in the FFT spectrogram. By utilizing this characteristic of the frequencies, we can distinguish between the open/close events of the door.

Next, we will look at how can we differentiate between the door events of multiple doors. Figure 1 shows a situation wherein a smartphone is placed in a room with two doors. The open event of Door1 shows that there is a gradually diminishing positive velocity component from the door toward the smartphone since the door starts moving from the closed position until the angle between the door frame and the door ($\theta 1$) becomes larger than that between the door frame and a line connecting the smartphone and the hinge ($\beta 1$), resulting in an increment in the observed frequency. From then, until the door reaches its opened position, the velocity component from the door toward the smartphone remains negative, resulting in a decrement in the observed frequency. When Door2 is opened, the velocity component from the door toward the smartphone remains positive all the way from the closed position to the opened position. This indicates that only an increment in the observed frequency can be expected. By utilizing this characteristic differences in the frequency shifts, we can distinguish between the door events of Door1 and Door2.

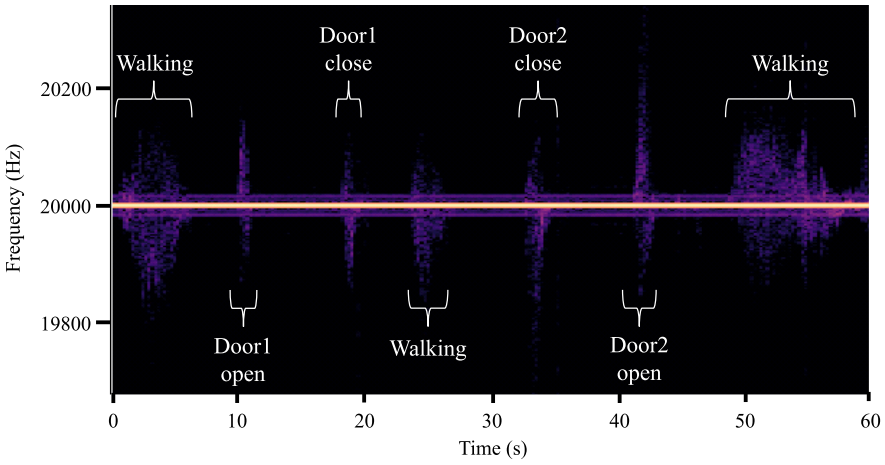

Fig. 4. FFT spectrogram of the front channel when the smartphone is placed 2.7 m from Door1 and 3.6 m from the Door2, as shown in Figure 3

3.2 Impact of Distance between Smartphone and Object

The open/close events of doors can be detected utilizing the Doppler shift, as shown above. We first investigated the effect of the distance between the smartphone and a door using the proposed method. We recorded acoustic signals emitted from the smartphone placed at various distances from the door. We used the Google Nexus 6P smartphone for active sound sensing. The smartphone emitted a sinusoidal sound wave with a frequency of 18 kHz from its speaker. The two inbuilt microphones (one in the front and the other at the back of the smartphone) recorded stereo sounds at a sampling rate of 44.1 kHz. Figure 2 shows the visualized FFT power spectra of the recorded sound signals that identify an open event when the phone was placed 1, 3, and 5 m from the door. Note that the front microphone and the speaker of the smartphone were positioned in the direction of the door. From these spectrograms, we can confirm that the frequency shift mostly occurred toward the positive direction during the open event of the door. In addition, the Doppler shift triggered by the door event could be observed by the smartphone even if the door was 5 m from it.

3.3 Time Variance of the Doppler Shift

Figure 3 shows the floor plan of our experimental environment, and Figure 4 shows an FFT power spectrum obtained when the open/close events of Door1 and Door2 occurred in the environment. During an open event for Door2, the door first moved toward the smartphone and then moved away from the smartphone. In this case, the time taken by the door to move toward the smartphone was relatively longer than the time taken by the door to move away from it. Therefore, we observed that the frequency shift caused by an open event first resulted in a strong shift toward the positive direction (higher than 20 kHz) and then resulted in a relatively weak shift toward the negative direction (lower than 20 kHz) of the FFT power spectrum. Conversely, during a close event of Door2, it moved toward the smartphone for a short time after it moved away from the smartphone for a longer time. This created a frequency shift in a direction opposite to that of an open event, creating a strong shift toward the negative side after creating a weak shift toward the positive side of the FFT power spectrum. In contrast, during the open/close event of Door1, we observed a short-duration frequency shift because Door1 (cabinet door) was smaller than Door2. Additionally, during an open event of Door1, we observed a frequency

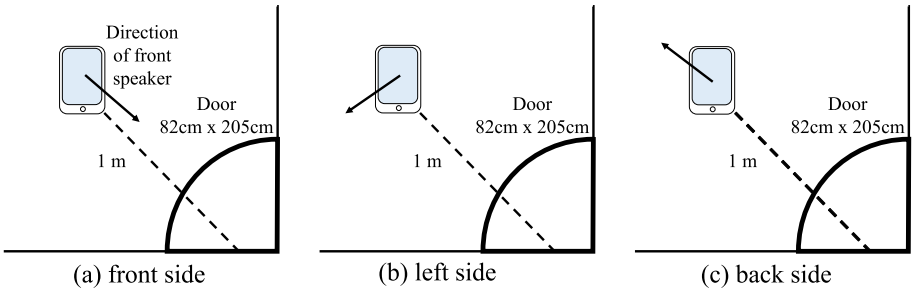

Fig. 5. Situations wherein a door is located (a) in front of the smartphone, (b) to the left-hand side of the smartphone, and (c) behind the smartphone. The smartphone emits a composite wave of two sine waves with frequencies of 18 and 20 kHz, respectively.

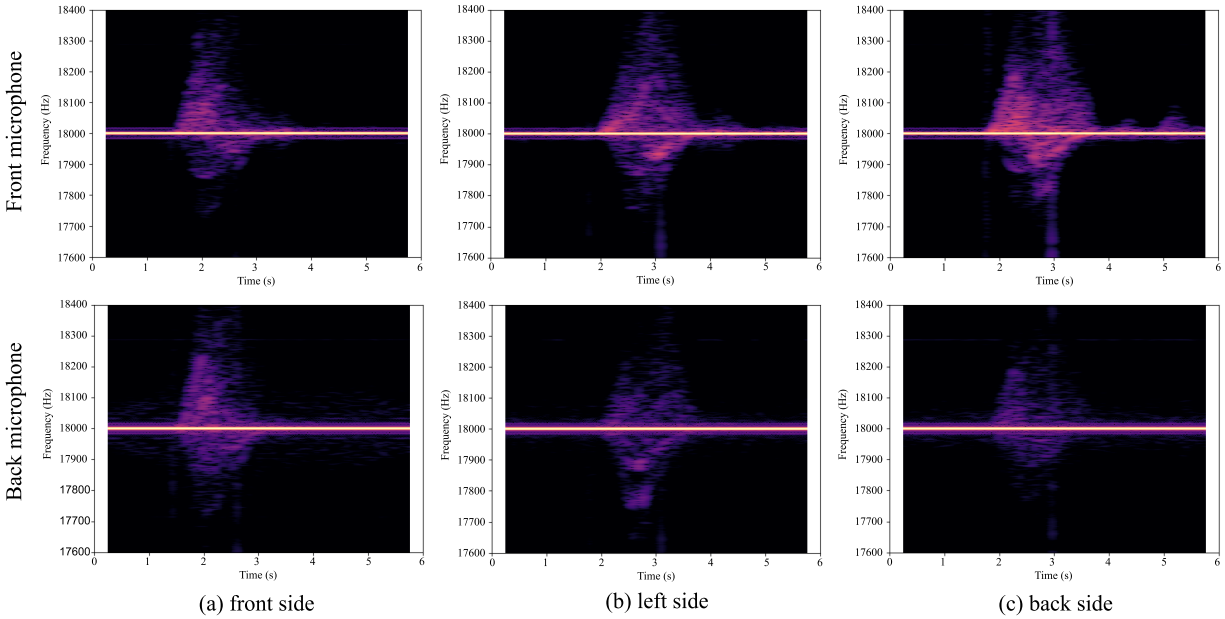

Fig. 6. Spectrograms of the door-opening events when the door is located (a) in front of the smartphone, (b) to the left-hand side of the smartphone, and (c) behind the smartphone. Spectrograms of the audio data from the front microphone are in the top row and the back microphone are in the bottom row.

shift that mainly comprised a shift toward the positive side. As above, because the observed frequency shift depends on the doors, it is necessary to capture the temporal patterns of the Doppler shift to recognize the open/close events.

Moreover, because the speed of walking is similar to that of the door event, the frequency shift caused by the walking events is also similar to that caused by the door event. However, since the duration of the walking event is usually longer than that of the door event, we believe that utilizing the temporal patterns of the frequency shift also enables to distinguish between the walking and door events.

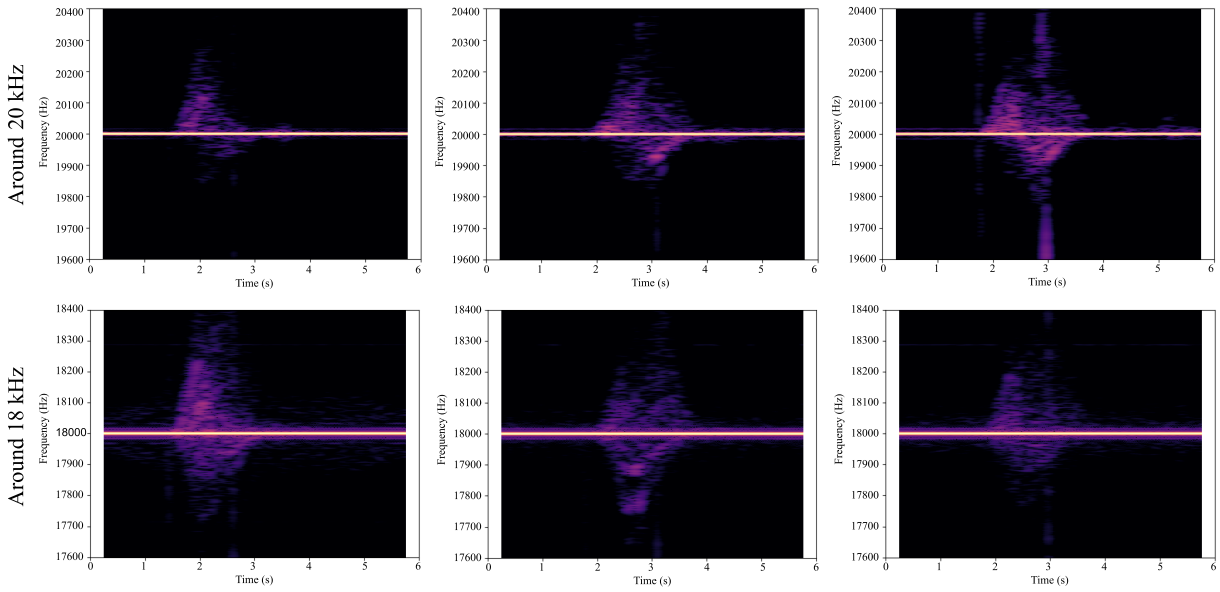

Fig. 7. Spectrograms of the audio data obtained by the front microphone when the door is located (a) in front of the smartphone, (b) to the left-hand side of the smartphone, and (c) behind the smartphone. The spectrograms in the top row show a frequency range of around 18 kHz, and the spectrograms in the bottom row show a frequency range of around 20 kHz.

3.4 Stereo Recording Using the Two Inbuilt Microphones

Figure 5 shows the three scenarios in which we collected sound data by varying the relative location of the door with respect to the smartphone. Figure 6 shows the FFT power spectrograms of an open event of the door for each situation. As shown in Figure 6, the sound features obtained from front and rear microphones of the smartphone were different. Even when the microphone did not face the door, the sounds of the microphone appeared to include a strong shift. This might have been caused by the reflected sounds from the surrounding walls.

3.5 Composite Signals Consisting of Two Sinusoidal Waves with Different Frequencies

The characteristics of the sound waves emitted from the speaker differed according to the frequency of the wave. For example, a lower frequency wave tended to have a higher amplitude than that of a higher frequency wave depending on the hardware properties of the speaker. In addition, sound waves with higher frequencies appeared to be more directional in comparison with sound waves with lower frequencies [24]. To obtain information from sound waves with both high and low frequencies, our method utilizes a composite wave. Figure 7 shows the spectrograms of the audio data obtained from the front microphone when a composite wave comprising 18- and 20-kHz sinusoidal waves was emitted when the smartphone was placed in three different configurations, as shown in Figure 5. As shown in the spectrograms of Figure 7, we could confirm that information obtained from the 20-kHz wave and that of the 18-kHz wave are different, indicating that utilizing the composite wave enables us to obtain information related to the direction of a door of interest.

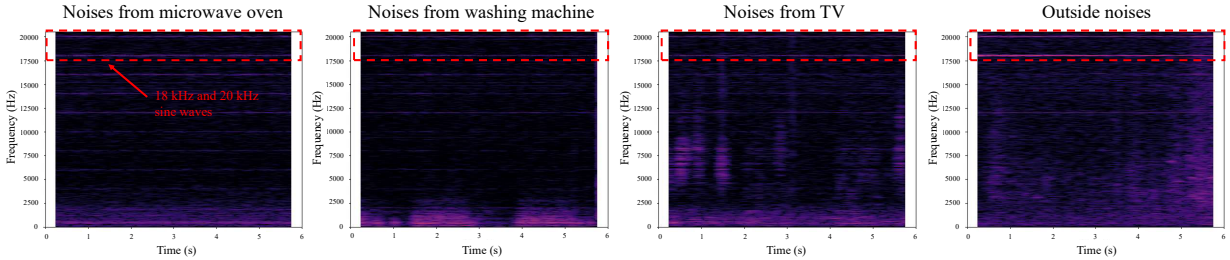

Fig. 8. Environmental noises recorded by a smartphone when a composite wave comprising 18- and 20-kHz sinusoidal waves was emitted. The distance between the smartphone and the microwave oven was approximately 3 m. The distance between the smartphone and the washing machine was approximately 3 m. The distance between the smartphone and the television was approximately 3 m. The distance between the smartphone and the entrance door was approximately 1 m.

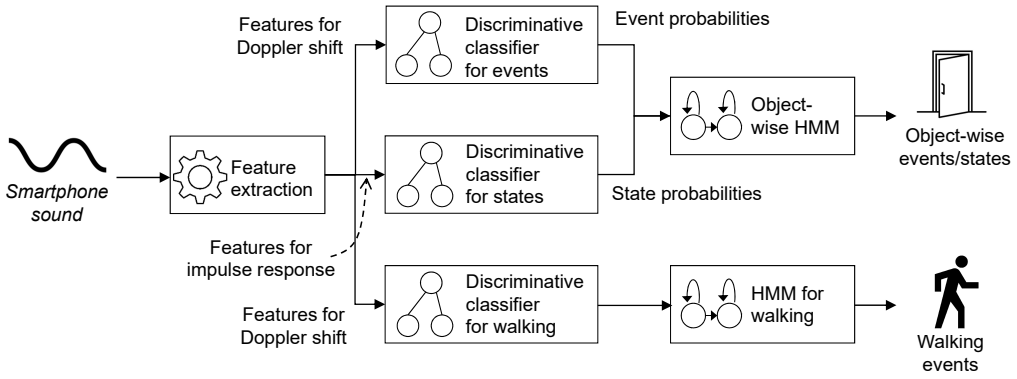

Fig. 9. Overview of the proposed recognition method

3.6 Daily Life Noises

Because our method focuses only on a high-frequency band, i.e., around 18- and 20-kHz, our method is robust against daily life noises. Figure 8 shows spectrograms of audio data containing daily life noises. As shown in the figure, the frequencies of noises caused by a microwave oven and washing machine are low. However, the television seems to suddenly emit high-frequency noises, which can deteriorate the performance of our method. While it is meaningless to emit inaudible sounds from televisions, we confirmed that the high-frequency noises are sometimes included in television sounds. In addition, when an entrance door is opened, outside noises are observed as shown in Figure 8, which can also deteriorate the recognition performance. To cope with these sporadic noises, our method is designed to contain generative models (HMMs), which is robust against such noises. Furthermore, we evaluated the performance of our method in such situations in Section 6.1.

4 METHOD

4.1 Overview

We assume that each door object has two events and two states, i.e., “open” and “close” events and “opened” and “closed” states. In addition, we regard that there are two events/states regarding to a person in an environment of interest; “walking” and “not-walking.” We recognize events/states of each object in the environment of interest

as well as the walking events for each time slice. In the proposed method, a smartphone emits a continuous composite sinusoidal wave and also emits occasional sine sweeps. We then use acoustic signals recorded from the front and back inbuilt microphones to recognize the events/states of the objects in the environment of interest. Figure 9 shows an overview of our recognition method. After extracting features, a feature vector sequence related to Doppler shift is fed into a discriminative classifier that recognizes the events of each object in the environment. In addition, a feature vector sequence related to the impulse response is fed into a discriminative classifier that recognizes states of each object. After that, the class probabilities from the discriminative classifiers are concatenated and fed into HMMs tailored to each object. As for the walking event recognition, the feature vector sequence related to Doppler shift is fed into a discriminative classifier for walking event detection. The class probabilities from the discriminative classifier are then fed into HMMs tailored for walking event detection.

As above, we apply the hybrid discriminative and generative approach for the event recognition. While discriminative classifiers are reported to outperform generative models, we attempt to deal with sporadic noises with the generative models, i.e., HMMs, which can capture the temporal regularity of the door/walking events in order to filter out the sporadic noises.

4.2 Sound Emission

Sine waves (with long durations) and sine sweeps (with short durations) were repeatedly emitted from the front speaker of the smartphone. A sine wave was emitted to obtain information about the Doppler shifts created by the moving objects. This can be utilized to detect door and walking events. A sine sweep was emitted to obtain acoustic information about the environment, which can be utilized to recognize the states of the objects.

The sine wave used in our method was a composite wave comprising two sine waves with frequencies of 18 and 20 kHz. These frequencies are not audible to the human ear. To generate the sine sweep, an excitation signal was emitted by the smartphone sweeping the frequencies from 18 to 20 kHz for 0.05 s. To avoid missing any events of the objects, the sine wave was continuously emitted for 10 s before emitting a sweep for 0.05 s. This process was repeated, and the timestamps of the emitted sweeps were saved in the smartphone.

4.3 Feature Extraction

We can extract the features related to the Doppler shift and impulse response from the acoustic signals obtained by both microphones.

4.3.1 Features Related to Doppler Shift. We extracted the features for event detection using the FFT power components around the 18- and 20-kHz frequencies. In the proposed method, we set up a 0.5-s sliding window with 96% overlap and applied a Blackman function to the window to calculate the FFT power spectrum of the audio data. We then extracted 4000 frequency bins from either side of the bandwidths of the 18- and 20-kHz sine waves. This produced a 16000-dimensional frequency vector sequence.

Next, we reduced the dimensionality of the vectors using time frame-wise averaging. Here, we used a 100-dimensional 50% overlapping sliding window along the frequency axis of each data frame, taking the average of the FFT power components within the window as one dimension. With this procedure, we can reduce the dimensionality of the vectors from 16000 to 320 dimensions while still retaining most of the characteristics of the original feature vectors.

4.3.2 Features Related to Impulse Response. We obtained the acoustic characteristics of the environment by analyzing the impulse responses of the emitted sine sweeps, as mentioned above. First, we calculated the FFT power spectrum of the audio data by applying a Blackman function to a 0.01-s sliding window with 85% overlap. This window size is considerably small and can be justified by the fact that the sweep length is 0.05 s.We used the

FFT power spectrum components corresponding to the frequency range of the sweep as features for recognizing the states of the objects.

Cowling et al. [7] revealed that the Mel-frequency cepstral coefficient (MFCC)-based feature extraction technique is one of the most appropriate approaches for environmental sound recognition systems, which has a recognition accuracy of 70%. Chen et al. [4] used MFCC features for recognizing bathroom activities, including showering, flushing, and urinating with high accuracy. Therefore, we decided to calculate 12-order MFCC features from the extracted FFT power spectrum components, which are also used as features.

4.4 Discriminative Classifier for Events

For each time slice, a feature vector consisting of features related to Doppler shift is fed into a discriminative classifier for events. The discriminative classifier is a random forest [2] that is used to compute class probabilities of open/close event classes of each object in the environment of interest and a class belonging to neither of above classes. Therefore, the random forest is a $(2N + 1)$-class classifier (this includes a class named "other" where no door event occurs), where N is the number of door objects in the environment. With this classifier, we can separate input signals containing information regarding events/states of multiple doors into a time-series of class probabilities for each class.

Here our method assumes that the smartphone periodically emits sine sweeps. Since the amplitudes of the sweeps are much larger than those of frequency shifts caused by the Doppler shift, the small frequency shifts might be ignored when the observed signals (or features extracted from the signals that include both the frequency shifts and sine sweeps) are directly fed into a generative model such as an HMM that learns a distribution of the input signals. In contrast, since a random forest classifier, which is the first-tier of our two-tier architecture, performs classifications based on thresholds, the scale of the input features does not affect the accuracy.

4.5 Discriminative Classifier for States

We prepared a discriminative classifier tailored to each door object that recognizes states of the objects using information obtained by sine sweeps. For each time slice, a feature vector consisting of features related to impulse response is fed into the discriminative classifier (random forest). As mentioned above, we repeatedly emitted sine waves and sine sweeps from the smartphone (after each 10-s sine wave signal, the smartphone emits 0.05-s sine sweep signals). Therefore, we can obtain the state information of the objects only during the considerably short period of time when a sine sweep is emitted. If we aim to gather training data without separating the instances where the sweep was emitted and where the sweep was not emitted, there is a strong possibility of incorrect and inefficient detection of the states of the objects. Since the time when the sine sweep signal is not emitted is much longer than the time when a sine sweep signal is emitted, a trained classifier can ignore the data corresponding to the time when the sweep is not emitted because the amount of the data is much smaller.

To solve this issue, we prepared a discriminative classifier for each door that separates the sweep information obtained when the door is in the opened state, the sweep information obtained when the door is in the closed state, the information obtained between the two sweeps when the door is in the opened state, the information obtained between two sweeps when the door is in the closed state, and the information that belongs to neither of the above states.

Therefore, the classifier is trained as a five-class classifier. Here we explain how to prepare a label for a feature vector at time t for the classifier. The feature vector is labeled as "opened@sweep" when the door was in the "opened" state and a sine sweep was emitted at time t. When the door was in the "closed" state and a sine sweep was emitted at time t, the feature vector is labeled as "sweep@closed." In contrast, when the door was opened and a sine sweep was not emitted at time t, the feature vector is just labeled as "opened." When the door was closed and a sine sweep was not emitted at time t, the feature vector is just labeled as "closed." Otherwise, the

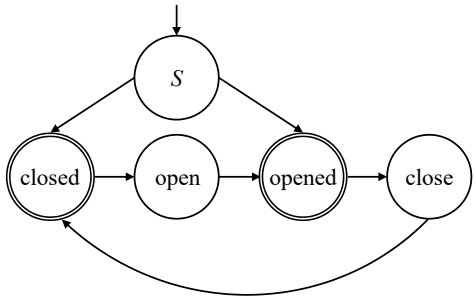

Fig. 10. State transition diagram of a handcrafted grammar describing state transitions across HMMs

feature vector is just labeled as "others." The outputs of the classifier are a class probability for each class for each time slice.

4.6 Object-wise HMM

We prepared a set of HMMs tailored to each object, consisting of a left-to-right HMM for each event/state of the object, i.e., "open" event, "close" event, "opened" state, and "closed" state. A sequence of class probability vectors obtained by the discriminative classifier for events of the object and a sequence of class probability vectors obtained by a discriminative classifier for states tailored for the object are concatenated and then fed into the HMMs. Therefore, the values of the observed variables of the HMMs correspond to the concatenated probability vectors, and output distributions of each state in the HMM are represented by using a Gaussian mixture model (GMM). The hidden state of the HMM shows the internal state of the event/state of the door. In the model, the hidden state at time t depends only on the previous hidden state at time $t-1$, i.e., a Markov process. Moreover, the observed variable at time t depends only on the hidden state at time t. Therefore, using the HMMs to decode the input sequence enables us to capture the temporal regularity of events/states. We used the Baum-Welch algorithm [22] to train the HMMs. When we decoded test data (probability vector sequence) by employing the trained HMM set, we used the Viterbi algorithm to estimate the most probable state sequence in/across the HMMs [22]. The recognition results show which HMM (event/state) a probability vector at time t corresponds to.

4.6.1 Decoding with Grammar. We prepared a left-to-right HMM for each event/state, and the Viterbi algorithm was used to find state transitions across the HMMs. This means that, when we decoded HMM inputs, we took into account a state transition from the last state of an arbitrary HMM to the first state of another HMM. This corresponds to, for example, a state transition from an "opened" HMM to a "close" HMM (and to all other HMMs of the door).

Here we restrict the state transitions among HMMs by employing a grammar handcrafted based on prior knowledge about state transitions of a door object. For instance, we can specify that a "close" event of a door occurs only when the door is in an "opened" state. We constructed such a grammar for door objects and Figure 10 shows the grammar represented in a state transition diagram. In Figure 10, the state "S" is the initial state and "closed" and "opened" are the final states. The state "S" is a dummy state introduced to limit the possible initial state of the door to be either "closed" or "opened."

After the state is transit to the "closed" state, for example, "open," "opened," "close," and "closed" events/states occur sequentially and are repeated. The state transition probabilities across the HMMs are determined based on the grammar (Figure 10). For example, the transition probabilities from the last state of the "opened" HMM to the

first states of all other HMMs except for the "close" HMM are set to be 0. In contrast, the transition probability to the "close" HMM is defined as 1. Moreover, we specify the possible final state to be either "opened" or "closed." When we decode the HMM inputs using the Viterbi algorithm, we refer to the grammar to obtain the transition probabilities among HMMs. Note that, because the Viterbi algorithm finds the most probable state transitions, it works even when the initial state of the door is not specified.

4.7 Discriminative Classifier for Walking Events

For each time slice, a feature vector consisting of features related to Doppler shift is fed into a discriminative classifier for walking events. The discriminative classifier is a random forest that is used to compute class probabilities of "walking" and "not-walking" classes in the environment of interest. Therefore, the random forest is a binary classifier.

4.8 HMMs for Walking Events

We prepared a set of HMMs, consisting of a left-to-right HMM for each walking event/state, i.e., "walking" and "not-walking." The training and decoding procedures are identical to those of the object-wise HMMs. Note that, because there is no restriction related to state transitions of the walking events, we assume that the transition probabilities between "walking" and "not-walking" are uniform.

5 EVALUATION

5.1 Data Set

We collected sensor data in real five environments. Figure 11 shows our five experimental environments and their settings. We installed Google Nexus 6P smartphone as shown in the figure. We recorded 44.1 kHz stereo audio using the front and back microphones of the smartphone. The smartphone emitted inaudible sinusoidal sound waves and sine sweeps during the experiment as described in the proposed method section. The smartphone also recorded time stamps when it emitted sine sweeps.

Environment 1 is a storage room with two doors and a cabinet. There are several discarded desktop PCs and shelves in the room. Environment 2 is a kitchen of a house with four doors. There are a refrigerator, a microwave oven, and other utensils. Environment 3 is a meeting room with one door. The distance between the door and the smartphone is approximately 5 meters. In this room, we verified the performance of our method when the distance between a door and the smartphone is long. There are four tables and 11 chairs in the room. Environment 4 is a room of a house with two doors including a sliding door. In this room, we verified the performance of our method when events/states of a sliding door are recognized. Environment 5 is a room of a house with three doors. Because Door1 in the room is an entrance door, outside noises were observed when the door was in the opened state.

In each environment, a participant conducted 8 sessions of data collection. To obtain ground truth, we recorded the sessions with a web camera. Each object has two events and two states, i.e., "open" and "close" events and "opened" and "closed" states. Throughout a session, the participant used all objects so that each event of the objects occurred twice in an arbitrary order. That is, in a session, for example, a door can be opened while a cabinet door can be closed. In another session, the door can be closed while the cabinet door is opened. As above, each object is used under different conditions in our experiment. Because the participant walked at random in the room and used the objects in a random sequence, the doors were used from different sides and in different positions. The duration of a session was approximately 240 seconds. The average duration of "walking" within a session was approximately 15.5 seconds. The "walking" durations entirely depend on the size of the environment and the preference of the user. For example, Environment 5 (Figure 11) is smaller than the other environments of interest. Furthermore, to simulate the natural activities of a user, the participant must exit the room while

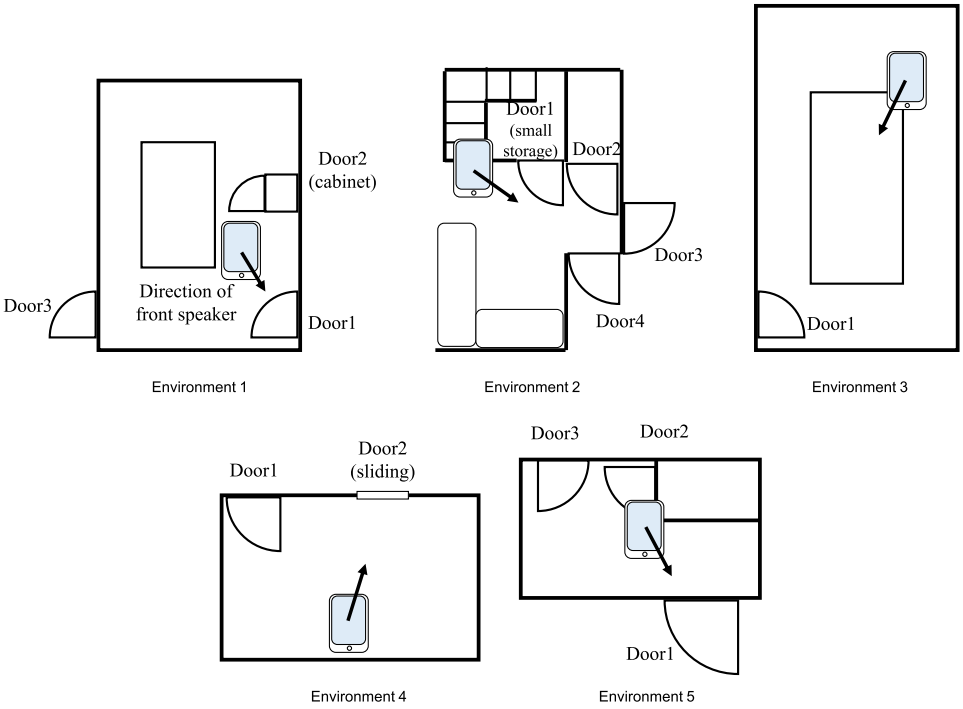

Fig. 11. Our experimental environments. The sizes of environments 1, 2, 3, 4, and 5 are 3.5 m × 4.4 m, 3.2 m × 4.4 m, 4.4 m × 6.8 m, 3.3 m × 2.2 m, and 2.6 m × 1.8 m, respectively.

closing the door behind him or her. These are the reasons why the average value derived from the "walking" data is only approximately 6% of the total duration of the session.

To investigate the effects of noises on the recognition performance, we collected additional three sessions of data in Environment 3 when another person was available (using a laptop PC). In addition, we collected additional three sessions of data in Environment 2 when a television was on. Furthermore, to investigate the effects of the ways of using doors on the recognition performance, we collected additional ten sessions of data in Environment 3 when each session was performed by different participant. Moreover, to investigate the effects of a newly installed object in an environment, we collected additional three sessions of data in Environment 2 after a chair was installed. These investigations will be described later.

5.2 Evaluation Methodology

We evaluated the performance of our method for each environment with leave-one-session-out cross validation. We prepared the following methods to investigate the effectiveness of the composite sine waves, two microphones, our two-tier architecture with discriminative classifiers, and a handcrafted grammar.

- *Proposed*: This is the proposed method.
- *w/o back MIC*: This method does not use sounds recorded from the back microphone in the feature extraction process. The other procedures are identical to those of *Proposed*.
- *w/o composite*: This method does not use composite sine waves. Instead, this method uses only 20 kHz sine waves. The other procedures are identical to those of *Proposed*.

Table 2. Classification accuracies for the proposed method in Environment 1

Object	Event/state	Precision	Recall	F-measure
Door1	open	0.542	0.886	0.672
	close	0.632	1	0.774
	opened	0.988	0.924	0.955
	closed	1	0.938	0.968
Door2	open	0.650	0.998	0.787
	close	0.59	0.935	0.723
	opened	0.987	0.940	0.963
	closed	1	0.899	0.947
Door3	open	0.474	1	0.643
	close	0.541	0.914	0.680
	opened	1	0.975	0.987
	closed	0.995	0.938	0.966

Table 3. Macro-averaged precision, recall, and F-measure for the proposed method in the five environments

Env.	Object	Precision	Recall	F-measure
1	Door1	0.790	0.937	0.842
	Door2	0.807	0.943	0.855
	Door3	0.752	0.957	0.819
2	Door1	0.838	0.963	0.884
	Door2	0.796	0.952	0.850
	Door3	0.834	0.972	0.886
	Door4	0.791	0.953	0.848
3	Door1	0.970	0.978	0.974
4	Door1	0.820	0.939	0.865
	Door2	0.681	0.738	0.705
5	Door1	0.844	0.975	0.895
	Door2	0.834	0.981	0.891
	Door3	0.869	0.976	0.912

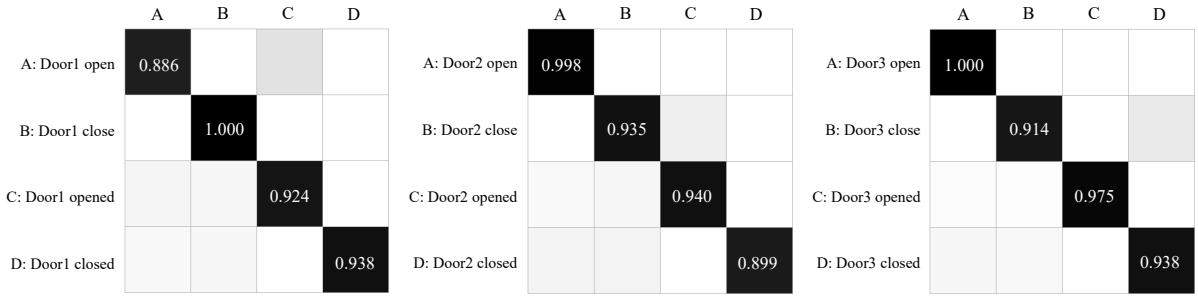

Fig. 12. Visual confusion matrices for the proposed method in Environment 1

- *w/o grammar*: This method does not use a handcrafted grammar for each object. That is, the transition probabilities from one HMM to the other HMMs are uniform. The other procedures are identical to those of *Proposed*.
- *HMM*: This method does not use discriminative classifiers for events/states. Extracted audio features for Doppler shift and impulse response are directly fed into an HMM set tailored for each object.

Recognition accuracy for each of the above methods is evaluated using the precision, recall, and F-measure, calculated based on the recognition results per window of data. The number of states of each HMM and the number of Gaussians in each state are 10 and 8, respectively.

5.3 Result of Door Event/State Recognition

5.3.1 Performance of the Proposed Method. Table 2 shows the classification accuracies for *Proposed* in Environment 1. In addition, Figure 12 shows visual confusion matrices for *Proposed* in Environment 1. As can be seen in the results, *Proposed* accurately recognized door events/states in Environment 1 even though the recorded sounds include noises caused by the participant (like Figure 4) because the participant walked around the room. However, the accuracies for the "open" and "close" events are somewhat poor compared to those of the "opened" and "closed" states. Figure 13 shows an example of the outputs of our method related to Door1 in Environment 1. As shown in the figure, the estimated start time of the "open" event has a small error (approximately 100

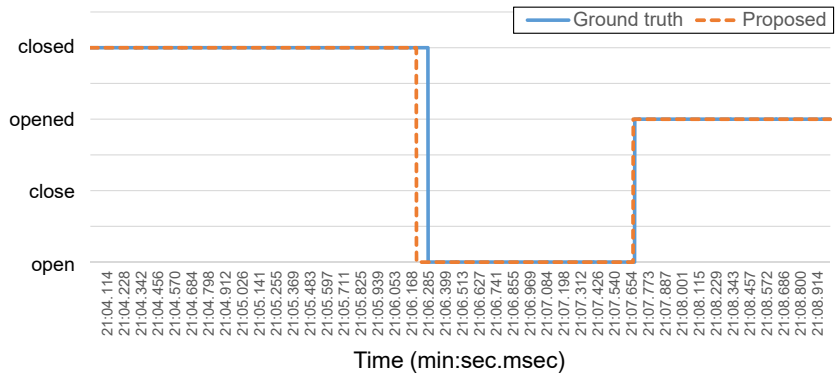

Fig. 13. Estimates of the proposed method for Door1 in Environment 1. The graph also shows the ground truth, which was obtained from video recordings.

msec). Since the number of instances (feature vectors) belonging to the "open" and "close" classes is small because the durations of these events are short, these small errors greatly affected the classification accuracies for these classes. However, because our method could capture the state changes of objects, the accuracies for the "opened" and "closed" states are almost perfect. We believe that most of the applications based on door event recognition can work well even when the detected starting/ending time of an event has a difference of 100 msec with the ground truth.

Table 3 shows the classification accuracies for *Proposed* in the five environments. Even though Door2 in Environment 1 is considerably smaller and located behind the smartphone, the accuracies for the door are just as good as the accuracies for the larger doors that are located in front of the smartphone. Similarly, as shown in the results of Door2 and Door3 in Environment 5, *Proposed* could precisely recognize events/states of doors behind the smartphone. In addition, even though the distance between the smartphone and Door1 in Environment 3 is about 5 m, *Proposed* could accurately recognize events/states of the door. In contrast, the accuracies for Door2 (sliding door) in Environment 4 are somewhat poor because it did not create a large enough Doppler shift to be captured by our smartphone. This is one limitation of the proposed method. However, we confirmed that the recorded sound sometimes captured weak frequency shifts when the sliding door was opened/closed even if the participant opened the door from outside the room. This is the reason why the accuracies for Door2 are not very poor.

5.3.2 Impact of Stereo Sound. Figures 14 - 18 show the macro-averaged F-measure of *w/o back MIC* for each door object in each environment. The accuracy of *w/o back MIC* for Door2 in Environment 1 is somewhat poorer than that of *Proposed*. This could be because the door is located behind the smartphone. In Environment 3, *Proposed* achieves a 17.2% higher accuracy than *w/o back MIC*. It can be assumed that the back microphone captured sound waves reflected by the wall behind the smartphone. In addition, because the back microphone seems to be more sensitive, which is used for noise cancellation, using the back microphone is also effective when recognizing events of objects located far from the smartphone. Interestingly, *w/o back MIC* achieves a 77.6% accuracy for the sliding door in Environment 4. In contrast, the F-measure for *Proposed* is 70.5%. Since the smartphone was very close to a wall as shown in Figure 11, we consider that the back microphone did not record any useful acoustic information because it did not capture sound waves reflected from the walls behind.

5.3.3 Impact of Composite Wave. Figures 14 - 18 also show the performance of *w/o composite*. The accuracies of *w/o composite* for Door1 in Environment 1 and Door 1 in Environment 3, which are located in front of the

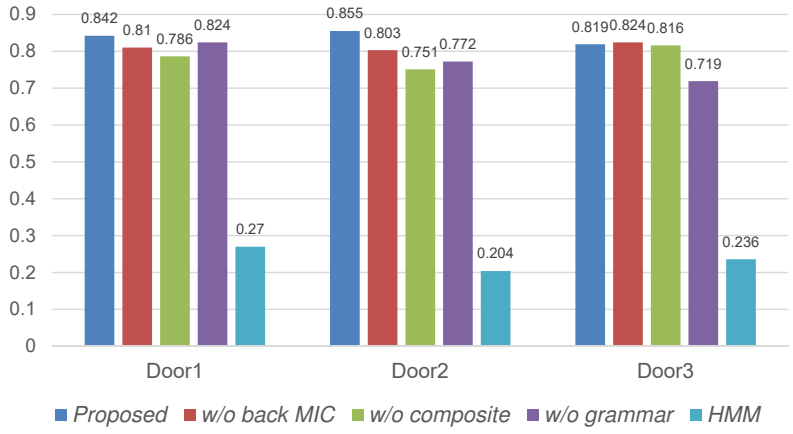

Fig. 14. Macro-averaged F-measures of the five methods in Environment 1

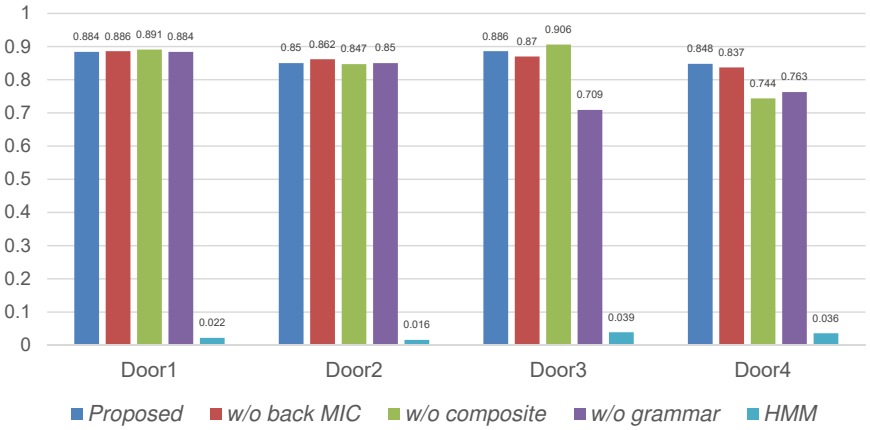

Fig. 15. Macro-averaged F-measures of the five methods in Environment 2

smartphone, are much poorer than those of *Proposed*. This could be because, when the smartphone faces the direction of a door, the door events do not cause strong frequency shift for 20 kHz signals as shown in Figure 7 (a), which can relate to the directionality of the high-frequency signals.

The accuracy of *w/o composite* for Door2 in Environment 4 is 38.4%. This is because we could not recognize any frequency shifts around a frequency range of 20 kHz when the sliding door was opened/closed. In contrast, we recognized small frequency shifts around a frequency range of 18 kHz.

5.3.4 Contribution of Grammar. Figures 14 - 18 also show the performance of *w/o grammar*. As can be seen in the results, using the grammar significantly improved the recognition accuracies of many objects. In particular, the accuracy for Door1 in Environment 3 improved by 30%. This could be because it is difficult to estimate states of a door located far from the smartphone. Figure 19 shows an example of the state transition of Door1 in Environment 1 estimated by *w/o grammar* in comparison to *Proposed*. As shown in the figure, *w/o grammar* outputs impossible state transitions, e.g., transition from the "open" event to the "opened" state. In contrast,

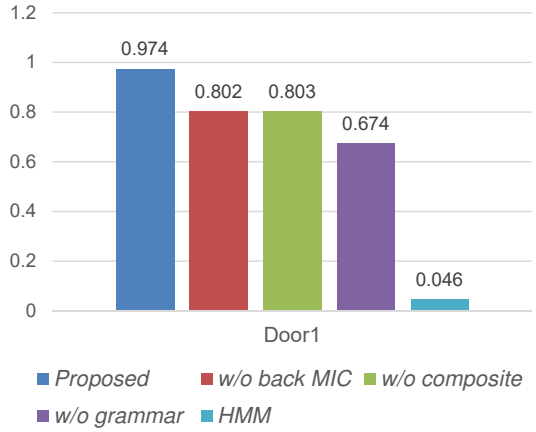

Fig. 16. Macro-averaged F-measures of the five methods in Environment 3

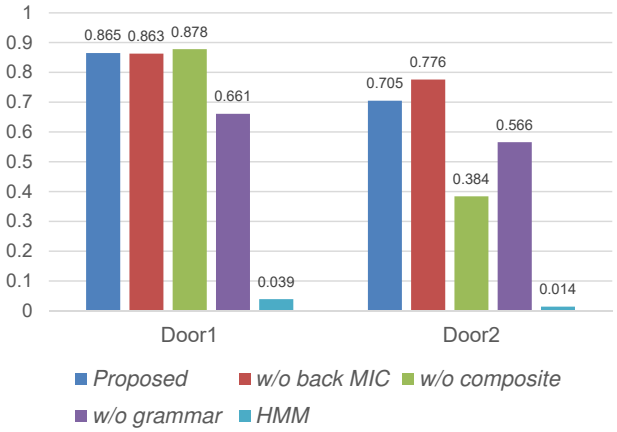

Fig. 17. Macro-averaged F-measures of the five methods in Environment 4

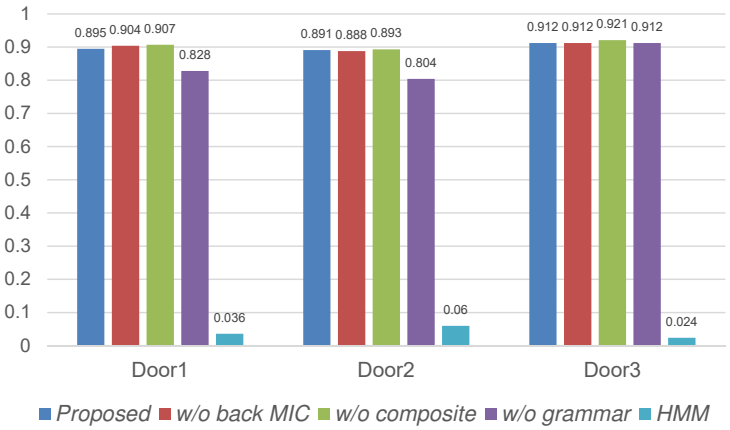

Fig. 18. Macro-averaged F-measures of the five methods in Environment 5

because *Proposed* decodes input signals based on the grammar, *Proposed* can output correct estimates even when it is difficult to estimate states of a door completely on the acoustic characteristics.

5.3.5 Contribution of Two-Tier Architecture. Figures 14 - 18 also show the performance of *HMM*. As shown in the results, it is impossible to recognize door events using the simple HMM architecture. As mentioned above, because the smartphone periodically emits sine sweeps whose amplitudes are much higher than those of frequency shifts caused by the Doppler shifts, the small frequency shifts are ignored when the extracted sound features are directly fed into the HMMs, which are generative models that learn distributions of observed variables. The accuracies in Environments 2, 3, 4, and 5 are extremely poor because almost all of the instances were mistakenly classified into the open/close classes. In contrast, our method, which is based on hybrid discriminative/generative models, is designed to detect the small frequency shifts using the discriminative classifier, which is robust against the problems regarding to the scale.

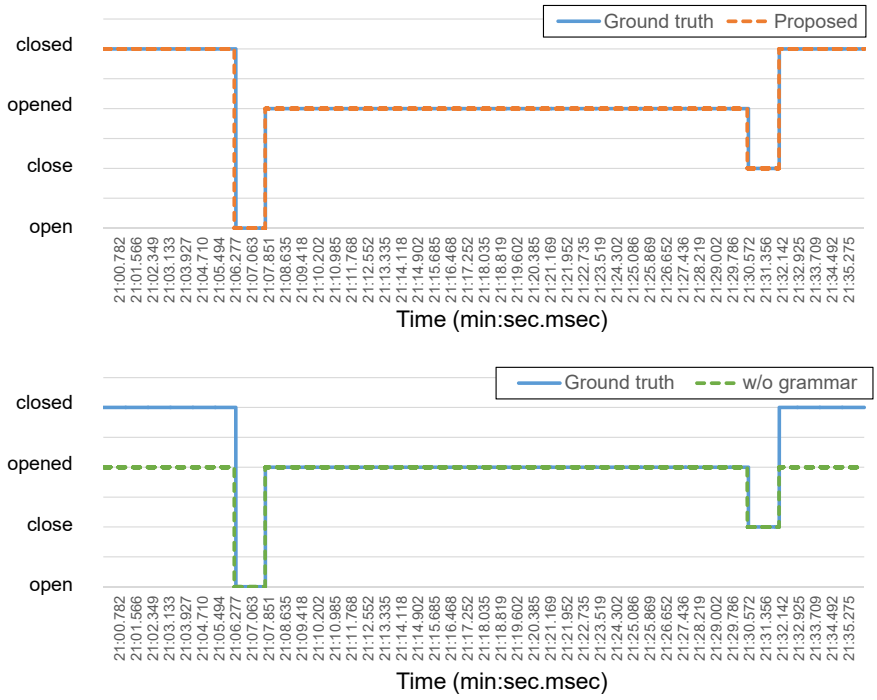

Fig. 19. Estimates of *Proposed* and *w/o grammar* for Door1 in Environment 1

Table 4. Accuracies of walking detection for the proposed method. (The results of Environment 3 are unavailable. Because there is only one door in the environment, the walking events did not occur.)

	event	precision	recall	F-measure
Env. 1	walking	0.637	0.876	0.738
	not-walking	0.976	0.910	0.942
Env. 2	walking	0.797	0.850	0.823
	not-walking	0.978	0.968	0.973
Env. 4	walking	0.887	0.814	0.849
	not-walking	0.941	0.966	0.953
Env. 5	walking	0.777	0.842	0.808
	not-walking	0.955	0.932	0.943

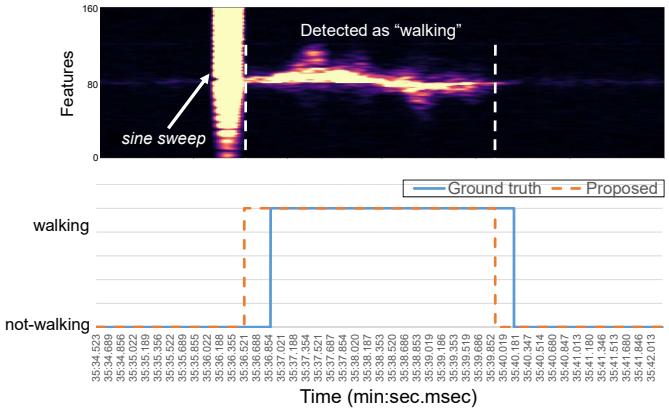

Fig. 20. Estimates of the proposed method (lower panel) and spectrogram related to frequencies used for classification features (upper panel)

Table 5. Accuracies of Door2 and Door3 in Environment 5 when the entrance door (Door1) was in the "opened" and "closed" states.

	Door	avg. precision	avg. recall	avg. F-measure
Door1 opened	Door2	0.833	0.983	0.891
	Door3	0.867	0.977	0.911
Door1 closed	Door2	0.836	0.980	0.891
	Door3	0.856	0.976	0.903

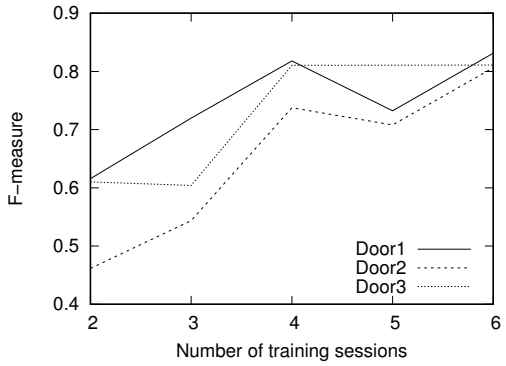

Fig. 21. Transitions in macro-averaged F-measure when the number of training sessions is varied

5.4 Results of Walking Event Recognition

Table 4 shows the recognition performance of the walking events. As shown in the results, our method could achieve highly accurate walking event detection. The accuracy for "walking" is somewhat poorer than that for "not-walking" because the durations of "walking" events are shorter than those of "not-walking" events. The proposed method failed to recognize the starting and ending parts of "walking" events as shown in Figure 20. In addition, because of the limited viewing angle of the video camera, it was difficult to precisely make the ground truth labels. The accuracies in Environment 1 were somewhat poorer than those in the other environments. This environment is a compact environment where Door3 opens into another very small storage space. When the Door3 was opened and the participant moved inside that storage space, we observed unexpected Doppler shifts caused by his movements even if he was not present in Environment 1, causing false positives in the classification. We believe that the reflected sound was emphasized by walls in the storage space since the storage space is very small.

6 DISCUSSION

6.1 Environmental Noises

We collected additional three sessions of data in Environment 2 when a television was on and a television program was played with medium volume. We trained a recognition model of *Proposed* on data collected when the television was off (8 sessions of data). The macro-averaged F-measures of Door1, Door2, Door3, and Door4 for the additional sessions were 0.885, 0.856, 0.715, and 0.798, respectively. As can be seen in the results, the F-measure of Door3 significantly decreased (see Table 3). A possible reason for this result is that the television programs in the authors' country contain audio frequencies up to 20 kHz (see Figure 8), which can interfere with

the composite sine wave emitted by the smartphone, hence worsening the recognition accuracy of the proposed method. We confirmed that the estimated starting/ending times of the "open" events of Door3 had large errors. However, the F-measures of the "opened" and "closed" states are still very high; 0.992 and 0.906, respectively.

As mentioned in Section 3.6, when an entrance door is in the opened state, outdoor noises can be observed by a smartphone. In Environment 5, we calculate the event recognition accuracies for Door2 and Door3 when Door1 (entrance door) was in the opened or closed state. When Door1 was in the opened state, outdoor noises were recorded as shown in Figure 8. The recognition accuracies for Door2 and Door3 were presented in Table 5. As shown in the results, even when Door1 was in the opened state, the proposed method, which consists of discriminative and generative models, achieved good recognition accuracies.

6.2 Effect of Additional Person

We collected additional three sessions of data in Environment 3 when a person was using a laptop PC, sitting between the smartphone and Door1. The macro-averaged precision, recall, and F-measure of these sessions for *Proposed* are 0.828, 0.974, and 0.883, respectively, when we train a recognition model on data collected when there was no person in the environment (8 sessions of data). The precision for the additional sessions was somewhat poorer than that of when there was no additional person in the environment as shown in Table 3 since the precisions for the open and close events are 0.670 and 0.643, respectively. This may be caused by small fluctuations of frequency caused by the person between the smartphone and Door1. However, the F-measures for the opened and closed states are as high as 0.974 and 0.979, respectively.

6.3 Way of Using Door

We collected eight sessions of data in Environment 3 when each session was performed by different participant to investigate the effects of the ways of using a door. The way of using the door was different from participant to participant. For example, a participant entered the room while opening the door. In contrast, another participant opened the door while standing still. We evaluate the additional data with leave-one-session out cross validation using *Proposed*. The macro-averaged precision, recall, and F-measure of Door1 were 0.935, 0.949, and 0.942, respectively. While the accuracies slightly decreased from those in Table 3, the F-measures for the door states are still high; 0.985 and 0.991 for the opened and closed states, respectively.

6.4 Additional Object

To investigate the effects of a newly installed object in an environment, we collected additional three sessions of data in Environment 2 after a chair was installed in front of Door3 and Door4. We trained a recognition model using *Proposed* on data collected before the installation (8 sessions of data). The macro-averaged F-measures of Door1, Door2, Door3, and Door4 for the additional sessions were 0.903, 0.857, 0.775, and 0.857, respectively. Since the chair was placed between the smartphone and Door3, the F-measure for Door3 dropped down to 0.775. We found that the output probabilities of the discriminative classifier for events after the chair was introduced were lower than the output probabilities for training data. Because the HMMs were trained on the output probabilities of the discriminative classifier for the training data, the trained HMMs could not deal well with the output probabilities for the test data.

To achieve robust recognition, we computed simulated pseudo output probabilities of the discriminative classifier for events and used them as additional training data for the HMMs. We simply reduced the actual output probabilities for the training data by multiplying each output probability by d (randomly selected from [0.3,0.9]). The macro-averaged F-measures of Door1, Door2, Door3, and Door4 when we used the robust method were 0.894, 0.847, 0.865, and 0.843, respectively. These results prove that, by using our robust method, we can reuse

training data collected under a given condition of an environment to achieve a high recognition accuracy when test data are collected after placing a small obstacle, i.e., chair, between the smartphone and a door.

6.5 Amount of Training Data

To investigate the amount of training data required to achieve accurate predictions using our recognition model, we performed the following test using eight sessions of data collected in Environment 1. We selected n sessions randomly as our training data and employed the remaining sessions as test data to obtain the recognition accuracy of the trained model. We repeated this process three times and calculated the average F-measure for each door. Then we performed the above test while changing the value of n from 2 to 6.

Figure 21 shows the relationship between the number of training sessions and the macro-averaged F-measure. As shown in the graph, we could achieve an average F-measure greater than 0.8 for Door1 (0.82) and Door3 (0.81) when we used 4 training sessions. However, the average F-measure for Door2 (0.74) is comparatively lower than the F-measures of Door1 and Door3 when 4 training sessions were used. The main reason for this is that the Door2 in Environment 1 is a cabinet door which is considerably smaller than Door1 and Door3. Therefore, the door events of Door2 created comparatively smaller Doppler shifts, hence making it harder for the recognition model to differentiate from Doppler shifts created by other movements inside the room using a small number of training sessions.

6.6 Structure of the System

In our method, we proposed to set up one smartphone per environment, where an environment represents a room in a house. Therefore, multiple smartphones have to be employed in order to implement this system for the entire house. In contrast to the distributed sensor method where sensors have to be mounted on every door or window of the house, the number of probes required by our method is equal to the number of rooms in the house. The reduced number of sensors is one of the main advantages of our method over the distributed sensor method. However, this raises the problem of how to communicate between these smartphones so that the entire network of the smartphones works coherently as a single system. This negatively affects the simplicity of our proposed method. Therefore, we consider each smartphone as a separate system, and each of them connects to Wi-Fi separately. While probing, audio data from each time slice are fed into the feature extraction process described in Section 4.3. Next, the extracted features will be sent to a server for event recognition via the Wi-Fi network.

6.7 Limitations

The evaluation revealed that it was difficult to accurately recognize events/states of a sliding door because our method is based on the Doppler shift created by the door movement. However, our experiments revealed that weak frequency shift was caused by events of the sliding door. In addition, the evaluation revealed that walking events in a room next to a room of interest were mistakenly detected. We believe that such false detection can be removed if another smartphone is installed in the next room.

7 CONCLUSION

This paper proposed a method to recognize door events via active sound probing using a smartphone. Our method achieved accurate recognitions employing a composite sine wave, stereo recording, a two-tier discriminative/generative recognition model, and a grammar that describes state transitions of a door. We collected data from different environments with different conditions and confirmed the effectiveness and validity of the proposed method.

At present, smart speakers such as Amazon Echo and Google Home have proliferated. As a part of our future work, we plan to implement our method on such a device to recognize events of the surrounding doors more accurately as they are equipped with more than two microphones. (Amazon Echo and Echo Dot are equipped with a seven-piece microphone array.)

ACKNOWLEDGMENTS

This work is partially supported by JST CREST JP-MJCR15E2, JSPS KAKENHI Grant Number JP16H06539 and JP17H04679.

REFERENCES

[1] Chris Beckmann, Sunny Consolvo, and Anthony LaMarca. 2004. Some assembly required: Supporting end-user sensor installation in domestic ubiquitous computing environments. In *International Conference on Ubiquitous Computing (UbiComp 2004)*. 107–124.

[2] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[3] Bradford Campbell and Prabal Dutta. 2014. An energy-harvesting sensor architecture and toolkit for building monitoring and event detection. In *the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*. 100–109.

[4] Jianfeng Chen, Alvin Harvey Kam, Jianmin Zhang, Ning Liu, and Louis Shue. 2005. Bathroom activity monitoring based on sound. In *Pervasive 2005*. 47–61.

[5] Brian Clarkson, Alex Pentland, and Kenji Mase. 2000. Recognizing user context via wearable sensors. In *Int'l Symp. on Wearable Computers (ISWC 2000)*. 69–75.

[6] Ionut Constandache, Xuan Bao, Martin Azizyan, and Romit Roy Choudhury. 2010. Did you see Bob?: human localization using mobile phones. In *MobiCom 2010*. 149–160.

[7] Michael Cowling. 2004. *Non-speech environmental sound recognition system for autonomous surveillance*. Ph.D. Dissertation. Griffith University.

[8] Biying Fu, Dinesh Vaithyalingam, Arjan Kuijper, Florian Kirchbuchner, and Andreas Braun. 2017. Exercise Monitoring On Consumer Smart Phones Using Ultrasonic Sensing. In *4th international Workshop on Sensor-based Activity Recognition and Interaction (iWOAR 17)*.

[9] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. 2012. Soundwave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1911–1914.

[10] Palanivel A Kodeswaran, Ravi Kokku, Sayandeep Sen, and Mudhakar Srivatsa. 2016. Idea: A system for efficient failure management in smart IoT environments. In *the 14th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys 2016)*. 43–56.

[11] Joseph Korpela, Ryosuke Miyaji, Takuya Maekawa, Kazunori Nozaki, and Hiroo Tamagawa. 2015. Evaluating tooth brushing performance with smartphone sound data. In *UbiComp 2015*. 109–120.

[12] Takuya Maekawa, Yasue Kishino, Yutaka Yanagisawa, and Yasushi Sakurai. 2012. Mimic sensors: Battery-shaped sensor node for detecting electrical events of handheld devices. In *Pervasive 2012*. 20–38.

[13] Takuya Maekawa, Yasue Kishino, Yutaka Yanagisawa, and Yasushi Sakurai. 2012. WristSense: wrist-worn sensor device with camera for daily activity recognition. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*. 510–512.

[14] Takuya Maekawa and Yuki Sakumichi. 2017. Easy to Install Indoor Positioning System that Parasitizes Home Lighting. In *European Conference on Ambient Intelligence*. Springer, 124–129.

[15] Takuya Maekawa, Yutaka Yanagisawa, Yasue Kishino, Katsuhiko Ishiguro, Koji Kamei, Yasushi Sakurai, and Takeshi Okadome. 2010. Object-based activity recognition with heterogeneous sensors on wrist. In *Pervasive 2010*. 246–264.

[16] Takuya Maekawa, Yutaka Yanagisawa, Yasushi Sakurai, Yasue Kishino, Koji Kamei, and Takeshi Okadome. 2009. Web Searching for Daily Living. In *SIGIR 2009*. 27–34.

[17] Takuya Maekawa, Yutaka Yanagisawa, Yasushi Sakurai, Yasue Kishino, Koji Kamei, and Takeshi Okadome. 2012. Context-aware Web search in ubiquitous sensor environment. *ACM Transactions on Internet Technology (ACM TOIT)* 11, 3 (2012), 12:1–12:23.

[18] Michael A Mahler, Qinghua Li, and Ang Li. 2017. SecureHouse: A home security system based on smartphone sensors. In *IEEE International Conference on Pervasive Computing and Communications (PerCom 2017)*. IEEE, 11–20.

[19] Kazuya Ohara, Takuya Maekawa, and Yasuyuki Matsushita. 2017. Detecting State Changes of Indoor Everyday Objects using Wi-Fi Channel State Information. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1, 3 (2017), 88.

[20] Shwetak N Patel, Matthew S Reynolds, and Gregory D Abowd. 2008. Detecting human movement by differential air pressure sensing in HVAC system ductwork: An exploration in infrastructure mediated sensing. In *International Conference on Pervasive Computing (Pervasive 2008)*. 1–18.

[21] Matthai Philipose, Kenneth P Fishkin, Mike Perkowitz, Donald J Patterson, Dieter Fox, Henry Kautz, and Dirk Hähnel. 2004. Inferring activities from interactions with objects. *IEEE Pervasive Computing* 3, 4 (2004), 50–57.
[22] Lawrence R Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 2 (1989), 257–286.
[23] Mirco Rossi, Julia Seiter, Oliver Amft, Seraina Buchmeier, and Gerhard Tröster. 2013. RoomSense: an indoor positioning system for smartphones using active sound probing. In *the 4th Augmented Human International Conference*. 89–95.
[24] Daniel A Russell, Joseph P Titlow, and Ya-Juan Bemmen. 1999. Acoustic monopoles, dipoles, and quadrupoles: An experiment revisited. *American Journal of Physics* 67, 8 (1999), 660–664.
[25] Shuyu Shi, Stephan Sigg, and Yusheng Ji. 2012. Passive detection of situations from ambient FM-radio signals. In *the 2012 ACM Conference on Ubiquitous Computing (UbiComp 2012)*. 1049–1053.
[26] Masaya Tachikawa, Takuya Maekawa, and Yasuyuki Matsushita. 2016. Predicting location semantics combining active and passive sensing with environment-independent classifier. In *UbiComp 2016*. 220–231.
[27] Emmanuel Munguia Tapia, Stephen S Intille, and Kent Larson. 2004. Activity recognition in the home using simple and ubiquitous sensors. In *International Conference on Pervasive Computing (Pervasive 2004)*. 158–175.
[28] Stephen P Tarzia, Peter A Dinda, Robert P Dick, and Gokhan Memik. 2011. Indoor localization without infrastructure using the acoustic background spectrum. In *MobiSys 2011*. 155–168.
[29] Yu-Chih Tung and Kang G Shin. 2015. EchoTag: accurate infrastructure-free indoor location tagging with smartphones. In *MobiCom 2015*. 525–536.
[30] Tim Van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse. 2008. Accurate activity recognition in a home setting. In *the 10th International Conference on Ubiquitous Computing (Ubicomp 2008)*. 1–9.
[31] Lloyd R Welch. 2003. Hidden Markov Models and the Baum-Welch Algorithm. *IEEE Information Theory Society Newsletter* 53, 4 (2003).
[32] Muchen Wu, Parth H Pathak, and Prasant Mohapatra. 2015. Monitoring building door events using barometer sensor in smartphones. In *the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015)*. 319–323.
[33] Zengbin Zhang, David Chu, Xiaomeng Chen, and Thomas Moscibroda. 2012. Swordfight: Enabling a new class of phone-to-phone action games on commodity phones. In *MobiSys 2012*. 1–14.

Received August 2018; accepted October 2018