Initial Investigation of a Notification System for Indoor Alarm Sounds Using a Neural Network

Takeru Kadokura[1], Kohei Watanabe[1], Yoshikaze Yanagiya[1], Elisa Sihombing[2], Syauqan Wafiqi[2], Yasuhiro Sudo[1], and Hiroshi Tanaka[1](✉)

[1] Kanagawa Institute of Technology,
1030 Shimo-ogino, Atsugi-shi, Kanagawa, Japan
{s1885004,s1521053,s1621092}@cce.kanagawa-it.ac.jp,
{sudo,h_tanaka}@ic.kanagawa-it.ac.jp

[2] Electronic Engineering Polytechnic Institute of Surabaya,
JL. Raya ITS – Kampas PENS Sukolilo, Surabaya 60111, Indonesia
elisasihombingspc@gmail.com, siconfix@gmail.com

Abstract. Many devices can inform the user of everything from a visitor's arrival at the door to a dangerous gas leak detection. For hearing-impaired people, there are some devices that can notify using light, etc., rather than by sound. However, these are individual devices and are relatively expensive due to their limited production volume. In this paper, a neural network was used as a method to classify alarm sounds of eight types of equipment. Two feature elements such as power spectrum and Mel Frequency Cepstrum Coefficients (MFCC) are taken as feature quantities to enter in this network, and its performance was evaluated. We implemented a neural network learned model on a Raspberry Pi and constructed a system that transmits classification results to a smartphone via Bluetooth. We generated 8 types of alarm sounds, plus indoor environmental sounds and speech sounds, for a total of ten kinds of sounds in the actual use environment of the classification experiment. This produced classification rates of 83.0% and 82.0% in experiments using learned models generated by power spectrum and MFCC. For the 8 alarm sounds, the classification rate was 87.5% by power spectrum and 77.5% by MFCC. It was confirmed that good performance could be obtained if power spectrum is used to determine feature elements in alarm sound classification.

Keywords: Alarm sound · Classification · Neural network · Feature element · Smart phone

1 Introduction

Quite a lot of equipment is available for use in the home to inform the owner of important events, ranging from visitors at the door to gas leaks, all of which require immediate attention, though with quite different responses and different urgency levels. Though the usual notification is a special sound signal, for hearing-impaired persons

© Springer Nature Switzerland AG 2019
S. Yamamoto and H. Mori (Eds.): HCII 2019, LNCS 11570, pp. 219–231, 2019.
https://doi.org/10.1007/978-3-030-22649-7_18

and those too distant to hear such signals, many products communicate by methods other than sound. However, remote warning devices are specialized in ways that make them much more expensive than ordinary devices [1, 2]. It would be useful to have a single system to recognize all these various alarm signals and communicate their messages through a single channel to their intended recipients, whether disabled, distant, or simply distracted.

Machine learning techniques have already been widely applied in fields such as image recognition, speech recognition [3] and automatic translation [4], so it seems reasonable to apply them to the classification of various alarm sounds, whether smoke alarms or kitchen trimers. Although there are studies aimed at detecting alarm sounds [5–7], it is two divisions of alarm sounds and non-alarm sounds, and a plurality of various alarm sounds are not classified. In addition, it is an examination of the element technology, and no investigation has been done on a system for communicating the occurrence of alarm sound to the user.

In this paper, we propose a system that recognizes various alarm sounds using machine learning, and transmits the notifications to the smartphone of an individual user. Here we describe the results of the primary prototyping of such a system.

2 Alarm Sound Classification

An image of the application of the proposed system is shown in Fig. 1. We propose a system which classifies various alarm sounds by machine learning and notifies the user by vibration of the user’s smartphone and displays the sound source on its screen. Although there are special devices that notify of an abnormality by other than sound, such as light, for users with hearing disabilities, they are all individual devices. As in the proposal shown in Fig. 1, it is thought that a useful system can be realized at low cost by detecting the sounds of all the alarms, identifying them and informing the user of the classification via smartphone, by means of vibration, etc.

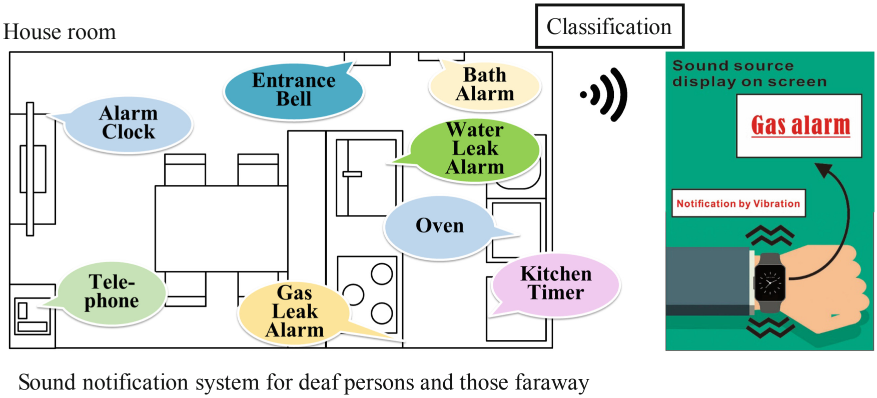


Fig. 1. Service image of proposed system

2.1 Feature Data Creation of Each Sound

Eight kinds of equipment, including a door alarm, two smoke alarms, a gas alarm, entrance bell, kettle alarm, and two timer alarms were selected as indoor alarm-sound producers in this investigation. The appearance of each alarm equipment and their spectrograms are shown in Fig. 2. The x-axis of the spectrogram graph is time (0 to 60 s), the y-axis is frequency (0 to 8000 Hz), and the sound level is −40 to 40 dB.

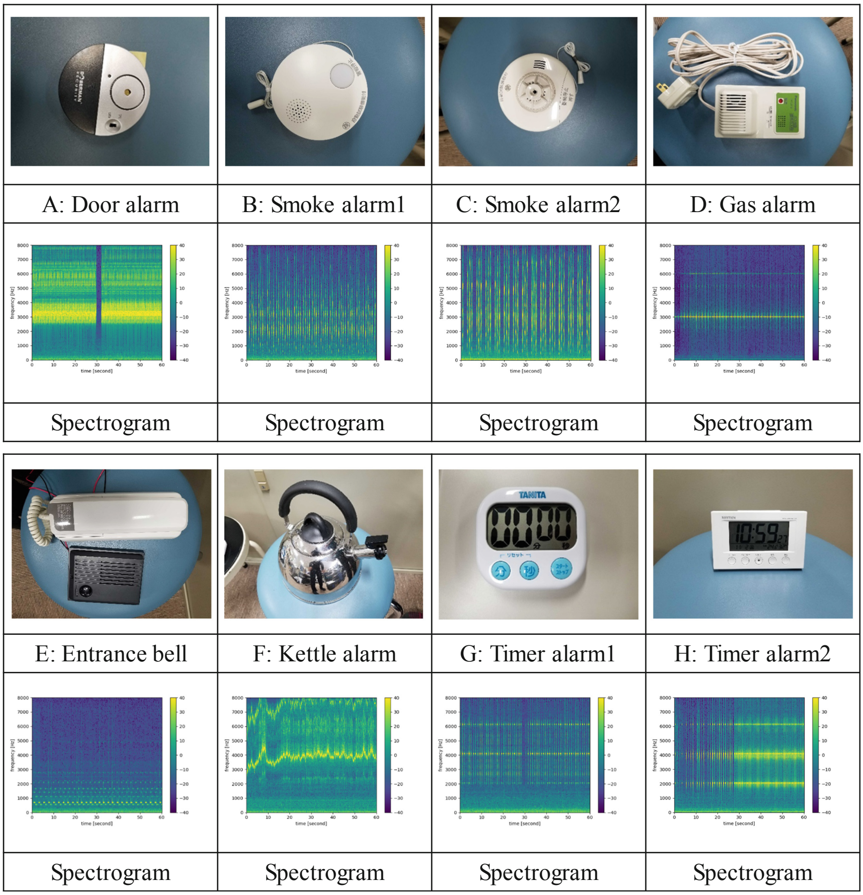
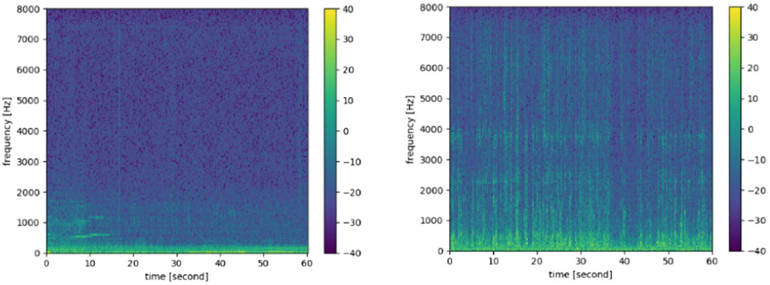


Fig. 2. Appearance of alarm equipment and their spectrograms

Here, it is necessary to consider the environment of any classification. That is, the classifier output produced even in the absence of an alarm sound. We selected two indoor sound environments: one in which air conditioners etc. are operating, but without conversation, i.e. just environmental sounds, and the other with normal speech sounds. These two sound backgrounds will be included for classification. The spectrograms of these two sound environments are shown in Fig. 3. It was confirmed that the spectrograms of each sound source (10 kinds of sound) were different and classification could be conducted by appropriate methods. To classify these sound sources, it was then necessary to select distinctive feature elements for each. In this investigation, we decided to use power spectrum and the Mel-Frequency Cepstrum Coefficients (MFCC) used in speech recognition as the feature elements for classification.



Environmental sound

Speech sound

Fig. 3. Spectrograms of environmental and speech sounds

Two data sets consisting of 60 s of a signal from each of the eight sound sources, plus environment and speech were acquired at a sampling rate of 16,000 Hz, and their power spectra up to 8 kHz was obtained for the extraction of feature elements. Each data set was divided into segments of 32 ms (512 samples), and the power spectrum was obtained by overlapping the bits every 10 ms (160 samples). Thus, 6000 pieces of power spectrum data were created for each data sample, for a total of 12,000 power spectrum data sets for use in the creation of the learned model described in Sect. 2.2. A hamming window was used to sample the power spectrum. Here, in order to eliminate the influence of the difference in sound volume due to differences in equipment and in distance, the result was normalized, that is divided by the maximum value.

The same data sets were used to extract MFCC as feature elements. Here, the number of Mel filter banks was 30, and 13-dimensional MFCC elements were extracted from 32 ms frame data samples using the Speech Signal Processing Toolkit (SPTK) [8]. The three components were then combined to generate 39 dimensional features, which were used to create a learned model.

2.2 Creation of Learned Model

The configuration of the neural network for learning is shown in Fig. 4. Each of the feature elements of the power spectrum and MFCC were used for training, both using the same neural network model, and a learned model was created. The intermediate layer is composed of two layers. A dropout configuration with 50% probability was introduced in the intermediate layer in order to avoid over-fitting.

Learning processes for the power spectrum and MFCC feature elements are shown in Fig. 5. The cross entropy error is used for the loss function, and the size of the mini

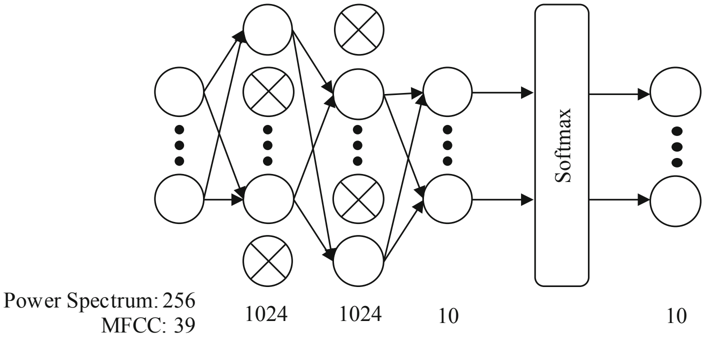


Fig. 4. Neural network configuration

batch is set to 20. We confirmed convergence of accuracy for the mini batch and loss obtained from the error function. For creation of a valid learned model, the number of epochs was set to 1000 to reach a stable status via this process. From the viewpoint of learning performance using feature data of 60 s, slightly better results were obtained when MFCC was used rather than power spectrum.

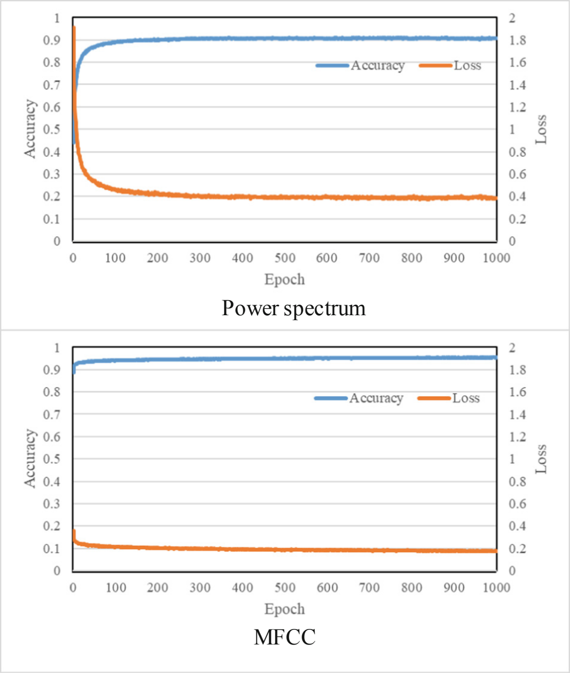


Fig. 5. Learning process

3 Classification Experiment

Giving the importance of early detection of an alarm sound, the classification performance for the data of the first 5 s of the alarm sounds was evaluated. Five seconds of data from each of the eight different alarm sounds, environmental and speech sounds were used to evaluate the classification performance by the learned network model obtained in Sect. 2. We generated sounds from prerecorded sound data from a Wav file and a speaker. Here, the mean and the standard deviation of the probability that is the output of the softmax function were examined together with the classification result for quality of classification.

3.1 Results by Power Spectrum

The classification performance when the power spectrum was used is shown in Table 1 as a confusion matrix. The overall accuracy was 96.0% (4 misjudgments among 100). Tables 2 and 3 show the average and the standard deviation of the output values of each classification target of the softmax function. Here, those whose value exceeds 0.1 are bolded. It can be understood from this average and standard deviation that misclassification occurred between the door alarm and kettle sounds as well as environmental and speech sounds.

Table 1. Result using power spectrum

	A	B	C	D	E	F	G	H	I	J
A	**9**	0	0	0	0	1	0	0	0	0
B	0	**10**	0	0	0	0	0	0	0	0
C	0	0	**10**	0	0	0	0	0	0	0
D	0	0	0	**10**	0	0	0	0	0	0
E	0	0	0	0	**10**	0	0	0	0	0
F	0	0	0	0	0	**10**	0	0	0	0
G	0	0	0	0	0	0	**10**	0	0	0
H	0	0	0	0	0	0	0	**10**	0	0
I	0	0	0	0	0	0	0	0	**8**	**2**
J	0	0	0	0	0	0	0	0	**1**	**9**

A: Door B: Smokel C: Smoke2 D: Gas E: Entrance
F: Kettle G: Timerl H: Timer2 I: Environment J: Speech

Table 2. Average values of softmax function for power spectrum

	A	B	C	D	E	F	G	H	I	J
A	**0.82**	0.02	0.02	0.00	0.00	0.11	0.01	0.01	0.00	0.00
B	0.01	**0.56**	**0.19**	0.05	0.00	0.00	0.01	**0.17**	0.00	0.00
C	0.01	0.04	**0.82**	0.02	0.00	0.01	0.01	0.07	0.00	0.01

(*continued*)

Table 2. (*continued*)

	A	B	C	D	E	F	G	H	I	J
D	0.01	0.02	**0.11**	**0.76**	0.00	0.00	0.01	0.09	0.00	0.00
E	0.00	0.00	0.00	0.00	**0.75**	0.00	**0.13**	0.00	0.03	0.09
F	0.00	0.00	0.09	0.00	0.01	**0.80**	0.04	0.00	0.04	0.03
G	0.00	0.01	0.01	0.00	0.02	0.02	**0.82**	0.01	0.01	**0.10**
H	0.00	0.08	**0.21**	0.03	0.00	0.00	0.01	**0.67**	0.00	0.01
I	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	**0.73**	**0.23**
J	0.00	0.01	0.01	0.00	0.10	0.01	**0.13**	0.00	**0.24**	**0.51**

A: Door B: Smokel C: Smoke2 D: Gas E: Entrance
F: Kettle G: Timerl H: Timer2 I: Environment J: Speech

Table 3. Standard deviation of softmax function for power spectrum

	A	B	C	D	E	F	G	H	I	J
A	**0.20**	0.04	0.02	0.00	0.00	**0.13**	0.02	0.01	0.00	0.00
B	0.00	0.03	0.03	0.01	0.00	0.00	0.01	0.04	0.00	0.00
C	0.00	0.01	0.03	0.01	0.00	0.01	0.01	0.01	0.00	0.01
D	0.00	0.02	0.06	**0.15**	0.00	0.01	0.02	0.08	0.00	0.01
E	0.00	0.00	0.00	0.00	**0.12**	0.00	0.08	0.00	0.03	0.05
F	0.00	0.00	0.07	0.00	0.01	**0.18**	0.04	0.00	0.05	0.03
G	0.00	0.01	0.01	0.00	0.01	0.02	0.06	0.01	0.01	0.03
H	0.00	0.07	0.03	0.01	0.00	0.00	0.01	0.07	0.00	0.01
I	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	**0.34**	**0.31**
J	0.00	0.01	0.01	0.00	0.04	0.01	0.08	0.00	**0.13**	**0.12**

A: Door B: Smoke1 C: Smoke2 D: Gas E: Entrance
F: Kettle G: Timerl H: Timer2 I: Environment J: Speech

3.2 Result by MFCC

Classification performance when MFCC is used is shown in Table 4. The overall recognition accuracy was 87.0% (13 misjudgments among 100). The erroneous classification of the environmental sound and the speech sound is similar to the result when using power spectrum, but erroneous classification also occurred for timer 1 and timer 2. The spectrogram in Fig. 2, shows that the very same sound is produced by each in the first time span. This is a very different result from the power spectrum case.

Tables 5 and 6 show the average value and the standard deviation of the output values of each classification target of the softmax function output, which are the same as in the power spectrum. In addition to environmental sound and speech sound, of course erroneous classifications of timer 1 and 2 occur.

Table 4. Result using MFCC

	A	B	C	D	E	F	G	H	I	J
A	**9**	0	0	0	0	**1**	0	0	0	0
B	0	**10**	0	0	0	0	0	0	0	0
C	0	0	**10**	0	0	0	0	0	0	0
D	0	0	0	**10**	0	0	0	0	0	0
E	0	0	0	0	**10**	0	0	0	0	0
F	0	0	0	0	0	**10**	0	0	0	0
G	0	0	0	0	0	0	**10**	0	0	0
H	0	0	0	0	0	0	**10**	0	0	0
I	0	0	0	0	0	0	0	0	**8**	**2**
J	0	0	0	0	0	0	0	0	0	**10**

A: Door B: Smokel C: Smoke2 D: Gas E: Entrance
F: Kettle G: Timerl H: Timer2 I: Environment J: Speech

Table 5. Average values of softmax function for MFCC

	A	B	C	D	E	F	G	H	I	J
A	**0.75**	0.00	0.09	0.03	0.00	0.10	0.00	0.02	0.00	0.01
B	0.00	**0.81**	0.05	0.01	0.01	0.00	0.08	0.03	0.00	0.00
C	0.00	0.02	**0.85**	0.03	0.01	0.01	0.01	0.05	0.00	0.00
D	0.00	0.04	0.04	**0.85**	0.00	0.00	0.05	0.02	0.00	0.00
E	0.00	0.02	0.02	0.01	**0.76**	0.00	**0.15**	0.00	0.00	0.04
F	0.00	0.00	0.02	0.00	0.00	**0.95**	0.01	0.00	0.00	0.02
G	0.00	0.09	0.05	0.04	0.01	0.00	**0.73**	0.03	0.00	0.04
II	0.00	0.08	0.00	0.03	0.07	0.00	**0.65**	**0.16**	0.00	0.00
I	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	**0.59**	**0.39**
J	0.00	0.03	0.00	0.02	0.02	0.00	0.01	0.00	0.05	**0.87**

A: Door B: Smoke1 C: Smoke2 D: Gas E: Entrance
F: Kettle G: Timer1 H: Timer2 I: Environment J: Speech

4 Development of Alarm Sound Notification System

4.1 System Requirements and Configuration

Based on the results in Sect. 3, we judged that a learned model capable of classifying each alarm sound had been created, and designed and developed an alarm sound notification system using this learned model. We constructed a system to notify a target user terminal such as a smartphone of an alarm and its source, as shown in Fig. 6. Vibration of the user terminal notifies the user of the occurrence of an alarm, and its display shows the source.

Since the classifier must always be in the power ON state, it is important for the classifier to have a low power consumption, so a Raspberry Pi with Bluetooth was used

Table 6. Standard deviation of softmax function for MFCC

	A	B	C	D	E	F	G	H	I	J
A	**0.20**	0.04	0.02	0.00	0.00	**0.13**	0.02	0.01	0.00	0.00
B	0.00	0.03	0.03	0.01	0.00	0.00	0.01	0.04	0.00	0.00
C	0.00	0.01	0.03	0.01	0.00	0.01	0.01	0.01	0.00	0.01
D	0.00	0.02	0.06	**0.15**	0.00	0.01	0.02	0.08	0.00	0.01
E	0.00	0.00	0.00	0.00	**0.12**	0.00	0.08	0.00	0.03	0.05
F	0.00	0.00	0.07	0.00	0.01	**0.18**	0.04	0.00	0.05	0.03
G	0.00	0.01	0.01	0.00	0.01	0.02	0.06	0.01	0.01	0.03
H	0.00	0.07	0.03	0.01	0.00	0.00	0.01	0.07	0.00	0.01
I	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	**0.34**	**0.31**
J	0.00	0.01	0.01	0.00	0.04	0.01	0.08	0.00	**0.13**	**0.12**

A: Door B: Smoke1 C: Smoke2 D: Gas E: Entrance
F: Kettle G: Timer1 H: Timer2 I: Environment J: Speech

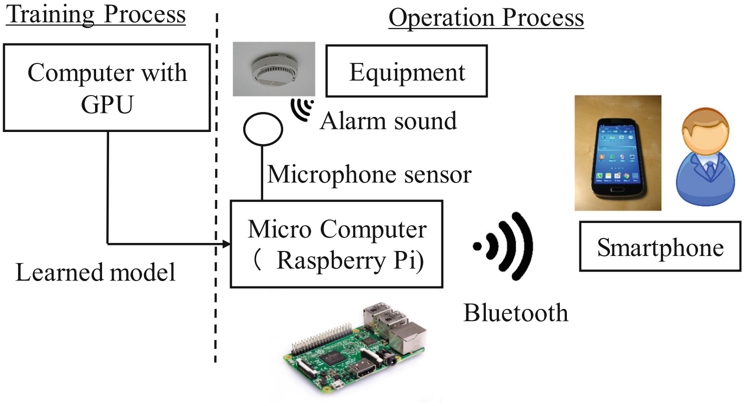


Fig. 6. Alarm notification system configuration

in this first prototype system. The learned model created in Sect. 3 was set up in the Raspberry Pi, which connects the microphone, and a smartphone was selected as the user terminal. The microphone is omnidirectional, and can be connected to the Raspberry Pi via USB, and is the same one used for data acquisition for the learning process.

4.2 System Design and Implementation

The system flowchart for activating the alarm sound is shown in Fig. 7. When an abnormality is detected, the alarm from the device continues to sound. It is necessary for the system to detect it and to notify the user as soon as possible. Therefore, in this prototype system, data acquired every 5 s was used, and the classification data was overlapped into 10 ms (duration 32 ms), that is, 500 pieces. Thus, 500 classification results are obtained, and the final result is determined by their majority decision. For

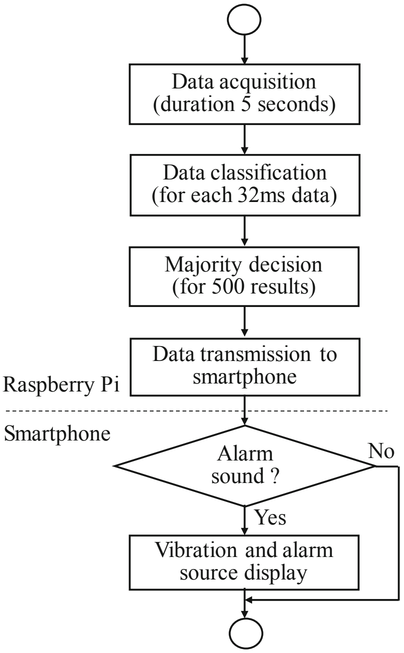

Data acquisition (duration 5 seconds)
Data classification (for each 32ms data)
Majority decision (for 500 results)
Data transmission to smartphone
Raspberry Pi
Smartphone
Alarm sound ?
No
Yes
Vibration and alarm source display

Fig. 7. System flowchart for alarm notification

each piece of alarm equipment, 50 s of sound were provided for 10-fold classification. Here, the identification result is transmitted to the smartphone using Bluetooth each time.

In the smartphone, when the classified alarm sound is received, i.e. the classification result of A, B, …, H, the body is vibrated and the source of the alarm sound is displayed on its screen. Environmental and speech sounds are reported as No Alarm.

The function of the smartphone that receives the classification result from Raspberry Pi by Bluetooth, makes its body vibrate, and displays the result, was implemented using MIT App Inventor 2 [8]. This development software provides functions to be implemented in a smartphone by dragging and dropping a visual representation of each instruction and function by using a graphical interface.

5 Initial System Evaluation

Initial evaluation of the alarm notification system composed in Sect. 4 was carried out. In this evaluation, the distance between the microphone sensor and each alarm sound source was 1 m, and the alarm sound to be classified was emitted continuously. Each 5 s of sound data from 8 types of alarm, and environmental and speech sounds were classified for evaluation.

An overview of the equipment used in the system is shown in Fig. 8. There are several of the sound-producing alarms, a microphone sensor, a Raspberry Pi, and a

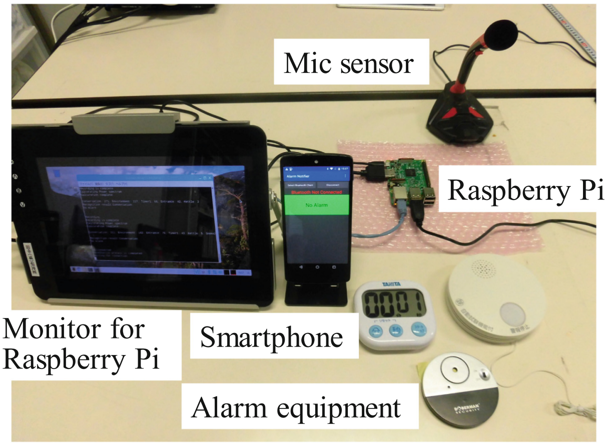


Fig. 8. Overview of components of alarm notification system

smartphone. The other device is a monitor for checking the output of the Raspberry Pi. As shown in this photo, very inexpensive equipment will suffice, so if a satisfactory level of classification performance can be secured there is a high possibility that a useful system can be realized for people with impaired hearing.

The classification experiments were carried out by implementing the learned model created by using power spectrum and MFCC in a Raspberry Pi. Results are shown in Tables 7 and 8 as a confusion matrix. There is little difference between the classification rates of 83.0% and 82.0%. In both cases, erroneous classification of timers 1 and 2 occurred. However, when MFCC was used, there was also misclassification of smoke 1 and the gas-alarm sound of smoke 2's alarm sound, while in power spectrum there was no additional error. For the 8 alarm sounds, the classification rate was 87.5% by power spectrum and 77.5% by MFCC. Although there were misclassification in

Table 7. Evaluation result by power spectrum

	A	B	C	D	E	F	G	H	I	J
A	**0.20**	0.04	0.02	0.00	0.00	**0.13**	0.02	0.01	0.00	0.00
B	0.00	0.03	0.03	0.01	0.00	0.00	0.01	0.04	0.00	0.00
C	0.00	0.01	0.03	0.01	0.00	0.01	0.01	0.01	0.00	0.01
D	0.00	0.02	0.06	**0.15**	0.00	0.01	0.02	0.08	0.00	0.01
E	0.00	0.00	0.00	0.00	**0.12**	0.00	0.08	0.00	0.03	0.05
F	0.00	0.00	0.07	0.00	0.01	**0.18**	0.04	0.00	0.05	0.03
G	0.00	0.01	0.01	0.00	0.01	0.02	0.06	0.01	0.01	0.03
H	0.00	0.07	0.03	0.01	0.00	0.00	0.01	0.07	0.00	0.01
I	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	**0.34**	**0.31**
J	0.00	0.01	0.01	0.00	0.04	0.01	0.08	0.00	**0.13**	**0.12**

A: Door B: Smokel C: Smoke2 D: Gas E: Entrance
F: Kettle G: Timerl H: Timer2 I: Environment J: Speech

Table 8. Evaluation result by MFCC

	A	B	C	D	E	F	G	H	I	J
A	**10**	0	0	0	0	0	0	0	0	0
B	0	**10**	0	0	0	0	0	0	0	0
C	0	**4**	**1**	**4**	0	0	**1**	0	0	0
D	0	0	0	**10**	0	0	0	0	0	0
E	0	0	0	**1**	**9**	0	0	0	0	0
F	0	0	0	0	0	**8**	0	0	0	**2**
G	0	0	0	0	0	0	**10**	0	0	0
H	0	0	0	**1**	0	0	**4**	**4**	0	**1**
I	0	0	0	0	0	0	0	0	**10**	0
J	0	0	0	0	0	0	0	0	0	**10**

A: Door B: Smoke1 C: Smoke2 D: Gas E: Entrance
F: Kettle G: Timerl H: Timer2 I: Environment J: Speech

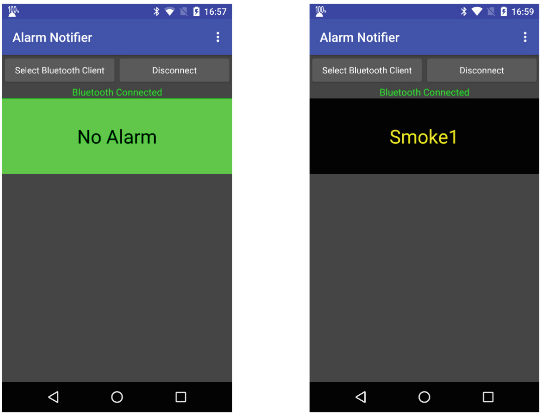


Fig. 9. Example of alarm display of smartphone

environmental and speech sounds when using power spectrum, this error is not a problem in terms of the required functions of the proposed system. From this result, it can be said that better performance can be obtained by using power spectrum when these particular alarm sounds are to be discriminated.

An example of the alarm display of the smartphone is shown in Fig. 9. Although there was classification error, it was confirmed that there was no problem in the operation of the system.

6 Conclusion

In this paper, we described a method for classifying various alarm sound sources and evaluated their classification performance using eight kinds of alarm sounds as well as conversational voice and environmental sounds. Then we proposed an alarm sound notification system using the created learned model, actually constructed the system as a prototype, and confirmed its basic functionality. We will conduct detailed evaluations and examine methods for enhancing accuracy in noisier environments. This time, the distance between the sound source and the microphone sensor was 1 m, and only two kinds of sound were involved: normal room sounds and speech sounds as the peripheral sound environment at the time of classification. In an actual life space, there will also be noises due to music, home appliance operation, etc. Improvement of classification performance and examination of evaluation methods considering various kinds of noises are future tasks.

References

1. Hearing Impairment Supplies Guide. https://moo-haya.ssl-lolipop.jp/nancho/kasai.html. Accessed 2 Jan 2019. (in Japanese)
2. “Warning light” informing people with hearing disorder. https://www.tbsradio.jp/13060. Accessed 26 Jan 2019. (in Japanese)
3. Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 5 p. (2013)
4. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: The International Conference on Learning Representations, pp. 1–15 (2015)
5. Carmel, D., Yeshurun, A., Moshe, Y.: Detection of alarm sounds in noisy environments. In: 25th European Signal Processing Conference (EUSIPCO), 5 p. (2017)
6. Raboshchuk, G., Nadeu, C., Jancovic, P., Lilja, A.: A knowledge-based approach to automatic detection of equipment alarm sounds in a neonatal intensive care unit environment. IEEE J. Transl. Eng. Health Med. **6**, 10 (2018)
7. Speech Signal Processing Toolkit (SPTK). http://sp-tk.sourceforge.net/. Accessed 26 Jan 2019
8. MIT App Inventor | Explore MIT App Inventor. http://appinventor.mit.edu/explore/. Accessed 26 Jan 2019