Alarm Sound Classification System in Smartphones for the Deaf and Hard-of-Hearing Using Deep Neural Networks

Yuhki Shiraishi
and Takuma Takeda

Faculty of Industrial Technology
Tsukuba University of Technology, Japan
Email: yuhkis@a.tsukuba-tech.ac.jp

Akihisa Shitara

Graduate School of Library, Information and Media Studies
University of Tsukuba, Japan
Email: theta-akihisa@digitalnature.slis.tsukuba.ac.jp

***Abstract*—For the deaf and hard-of-hearing to be able to go out safely, they must be able to recognize alarm sounds (horns, bicycle bells, ambulance sirens, etc.) among various environmental sounds. Therefore, it is crucial to be able to transmit these kinds of sounds to such people, even in noisy environmental conditions. In this paper, we propose and develop an alarm sound classification system using deep neural networks. The system works on smartphones that can always be carried by the users when they are going out. Besides, we performed evaluation experiments to verify the effectiveness of the system using the 5-fold cross-validation method. Furthermore, we evaluate the classification rate for unlearned data and re-evaluate one by adding data downloaded from the web. We also discuss the limitations of the system to improve it and make it more useful.**

Keywords–Alarm sound; Classification; Deaf and hard-of-hearing; Neural network; Smartphone.

I. INTRODUCTION

Over 5% of the world's population (466 million people) has disabling hearing loss as stated in [1]. In order for these people to be able to go out safely, they must be able to recognize alarm sounds (horns, bicycle bells, ambulance sirens, etc.) directly linked to a safe and secure life, among various environmental sounds. Therefore, there is need for a system that distinguishes these specific alarm sounds from environmental sounds and transmits them to those with disabling hearing loss.

In recent years, Deep Neural Networks (DNNs) have been attracting attention; DNNs automatically learn alarm sounds to be recognized, and they automatically acquire the features of these sounds. With DNNs, high-precision classification is expected even when the sound quality is affected because of the movement of objects or noisy environments.

In this research, we develop an alarm sound classification system using DNN (Figure 1). As a result, hearing-impaired people will be able to recognize alarm sounds and go out safely. Our aim is to build a system that uses a smartphone because the users carry smartphones when they go out.

In this paper, we propose an alarm sound classification system using DNN and confirm essential classification performance. Moreover, we develop a smartphone application that recognizes the siren of an ambulance, the bell of a bicycle, etc., and sends it to the user. We perform evaluation experiments to verify the effectiveness of the system using the 5-fold Cross-Validation (CV) method. Furthermore, we evaluate the classification rate for unlearned data and re-evaluate one by adding data downloaded from the web. We also discuss the limitations of the system to improve it and make it more useful in the future.

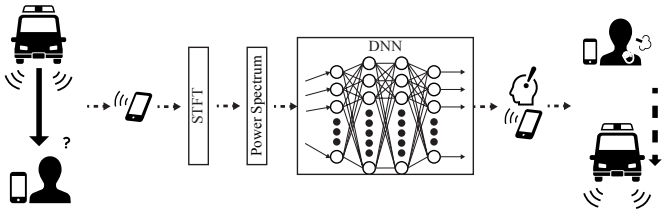

Figure 1. Alarm sound classification and transmission systems.

II. RELATED WORK

Ontenna [2] is an interface that focuses on vibration, which lets the user recognize sounds by real-time vibration. The system was created to recognize sound, so there is no system to tell the user the type and direction of the sound. However, the Ontenna system is tiny and lightweight. In the system, a sound of 0–90 dB was converted into 256 steps of vibration and light intensity. The sound feature is transmitted to the user through some kinds of vibration.

Google Live Transcribe [3] is mainly for voice recognition, but can also recognize environmental sounds. The only alert sound supported by the system is the horn of the car. Moreover, since the main feature of the system is voice recognition, there is no ability to communicate with the user via a vibration or through pop-up notifications.

Wavio SeeSound [4] can send sounds to the user via vibration and pop-up notifications. However, the system works indoors and does not support outdoor use.

Takeda et al. [5] proposed a system for classifying alarm sounds using a multilayer perceptron neural network. However, the alarm system only targets straightforward beep sounds in oxygen concentrators.

Nicholas et al. [6] present the first mobile audio sensing framework built from coupled deep neural networks that simultaneously perform everyday audio sensing tasks. However, the target sounds are from diverse acoustic environments such as bedrooms, vehicles, or cafes. The classification ratio is at most about 90%, which is inadequate for safety alarm recognition.

 ISBN: 978-1-61208-761-0

Meanwhile, Jain et al. [7] examine how Deaf and Hard-of-Hearing (DHH) people think about sounds in the home, and they explore potential concerns. Findlater et al. [8] conducted an online survey with 201 DHH participants to investigate preferences for mobile and wearable sound awareness systems. The reviewed studies support the importance of alarm sound classification systems.

III. Development System

In this system, the classification and transmission application run on a smartphone without internet connection. Since users of this system are DHH or people with disabling hearing loss, a non-sound notification system is required. Therefore, the developed system displays the names of alarm sounds on the screen when such sounds occur.

The basic flow of the proposed system is as follows:

1) Collect environmental sounds with a smartphone.
2) Notify smartphone when an alarm sound is identified.

Deep learning is used as a classification method. To create learning data, we collected sound data such as ambulance sirens, horns, and bells, to be classified and transmitted. We pre-collected these sounds in a real environment using smartphones. The reason why we collected the data in the real environment instead of using the pure tone of the warning sound is to make full use of the generalization ability of deep learning.

We performed data reduction and data screening on the alarm sound data collected in various environments, and we created a learning database.

Keras [9] was used for implementing deep learning algorithms. Keras was a wrapper library for Tensorflow [10], and now Keras is officially integrated into Tensorflow. Besides, it supports not only Linux servers but also Android and iOS, which makes possible application development more straightforward.

Figure 2 shows a snapshot of the ongoing developed application. Presently, the application works only on the iPhone, which is programmed using the Swift programming language. By using Apple's neural network library, we can import the learned weight data using Keras to iPhone.

IV. Classification Algorithm

The alarm classifying flow consists of the following three steps.

1) Continuous collection of environmental sounds.
2) If volume data exceeding the threshold is detected, record audio data for a certain period.
3) Specify the alarm class (horns, bicycle bells, ambulance sirens, etc.) of the recorded audio data.

Besides, because the nature of the alarm sound tends to be monotonous, we apply the Short-Time Fourier Transform (STFT)

$$STFT(t,\omega) = \int_{-\infty}^{\infty} x(\tau)h(\tau - t)e^{-j\omega\tau}d\tau, \qquad (1)$$

where $x(t)$ is sound data, and $h(t)$ is a window function to the sound data collected by the above threshold processing.

After STFT, the power spectrum of STFT is converted to the log scale, which is used as an input to the DNN. Finally,

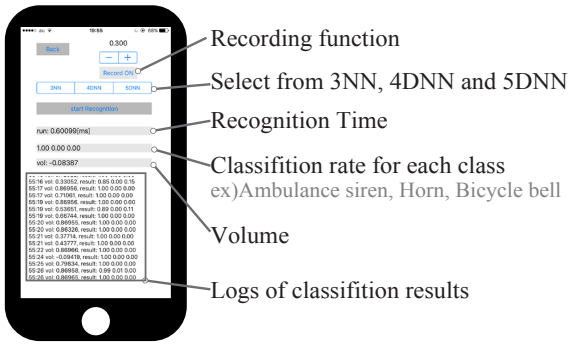

Figure 2. Snapshot of develop smartphone application.

real-time classification is performed by applying an integrated process to the one-time classification results that DNN has repeatedly determined for all audio data.

ReLU (2) is used for the activation function, Softmax cross entropy (3) is used for the error function, and Adam [11] is used for the learning algorithm, where t_k is the correct label (one-hot expression), and y_k indicates the network output.

$$f(u) = max(u, 0) \qquad (2)$$

$$E = -\sum_{k} t_k \log y_k \qquad (3)$$

The operation of the classification application is as follows:

1) Use a smartphone microphone and collect sound every 1024 frames using 32-bit single-precision floating-point numbers (-1.0 to 1.0).
2) When the absolute value of the buffered single-precision floating-point buffer exceeds the threshold value (0.3), identification processing starts.
3) Multiply the buffer by 2^{31} and change the buffer range to a 32-bit integer type, then execute STFT.
4) Input of logarithmic power spectrum to DNN.
5) Display the classification result on the screen.

In a real environment, the target sound would continue to resonate so that the classification result would be displayed multiple times for one occurrence of the target sound. Therefore, considering the importance of desired alarm sounds, the final classification result is determined by the following algorithm (called integrated judgment process). As a result, the classification ratio and reliability are expected to be improved.

1) Evaluate sounds continuously (more than once to less than ten times).
2) If there is more than one classification result from a specific sound other than noise,
   a) Calculate the sum of outputs.
   b) The largest of the noise exclusions is used as the final classification result.
3) If all classification results are noise,
   a) Regard the final classification result as noise.

Figure 1 shows the flow of the entire operation up to the classification result determination.

 ISBN: 978-1-61208-761-0

TABLE I. 5-FOLD CV FOR 5 TYPES OF ALARMS.

Number of layers	*Classification rate*
3	0.9845
4	0.9867
5	0.9924

TABLE II. CLASSIFICATION RESULTS IN A NOISY ENVIRONMENT (BEFORE APPLYING THE INTEGRATED JUDGMENT PROCESS).

	TP	*FP*	*FN*	*TN*	*Prec.*	*Recall*	*F-value*	*Max vol[dB]*
Horn	545	0	87	2232	1.00	0.86	0.92	98.1
Bicycle bell	502	0	113	2249	1.00	0.81	0.98	127.7
Ambulance	572	1	56	2336	0.99	0.91	0.95	90.0
Fire alarm	631	1	57	2176	0.99	0.91	0.95	93.2
Noise	298	262	2	2563	0.53	0.99	0.69	100.3

V. EXPERIMENTS

A. Basic performance of the classification system

In addition to the two types of manually collected sound data (ambulance sirens and bicycle bells), we downloaded a total of 18 horn sound data from the web page [12]. We also manually recorded fire alarm sounds during evacuation drills. Furthermore, we added a noise class to handle cases where sounds other than the target sounds are generated. We collected six types of noises: footsteps, car driving sounds, voices, door opening/closing sounds, hitting desks, and rubbing plastic bags.

Training and evaluation were performed on 3-layer NN, 4-layer DNN, and 5-layer DNN. We performed STFT with 1024 frame for the 44.1 kHz 32 bit sound. We carried out a 5-fold CV for 25 000 pieces of training and evaluation data (5000 pieces $\times$ 5 classes) with a maximum of 1000 epochs (input layer: 513, hidden layer: 128, output layer: 5).

A 5-fold CV is described as follows. First, we divide all data into five groups. Next, data from one group are used for the test and the data from the other four groups are used for the learning. Finally, the learning process is repeated five times by using five different test groups.

Table I shows the experimental results. The classification results in the table are above 98% for all NN/DNNs. In the following experiments, we used the five-layer DNN because it gives the highest classification rate.

B. Performance in a noisy environment

Next, the experiment was performed in a noisy environment of 50.5 to 100.3 dB. In this study, we assumed that the noise originating from outdoors was mainly the noise of cars, and repeatedly evaluated the noise from driving cars 100 times (after applying the integrated judgment process). At that time, we recorded the maximum volume of each target sound.

Table II shows the classification results before applying the integrated judgment process, and Table III shows the results after applying the integrated judgment process. After applying the judgment process, it was possible to classify these sounds with an average F-measure of more than 99% in a real environment.

C. Performance for unlearned horn sounds

In this algorithm, we performed feature extraction and identification using STFT. In particular, since the quality of horn sounds differs depending on the type, the frequency characteristics also differ. Therefore, there is concern that generalizability of the performance of new types of horn sounds is perhaps low.

TABLE III. CLASSIFICATION RESULTS IN A NOISY ENVIRONMENT (AFTER APPLYING THE INTEGRATED JUDGMENT PROCESS).

	TP	*FP*	*FN*	*TN*	*Prec.*	*Recall*	*F-value*	*Max vol[dB]*
Horn	100	0	0	400	1.00	1.00	1.00	98.1
Bicycle bell	100	0	0	400	1.00	1.00	1.00	127.7
Ambulance	100	0	1	400	1.00	0.99	0.99	90.0
Fire alarm	100	0	1	400	1.00	0.99	0.99	93.2
Noise	100	2	0	398	0.99	1.00	0.99	100.3

TABLE IV. UNLEARNED HORN CLASSIFICATION (BEFORE APPLYING THE INTEGRATED JUDGMENT PROCESS).

	TP	*FN*	*Classification rate*
Horn 1	62	16	0.79
Horn 2	43	11	0.80
Horn 3	42	8	0.84
Horn 4	55	23	0.71
Horn 5	67	8	0.89
Horn 6	36	29	0.55
Horn 7	50	33	0.60

Therefore, we examined the classification in the case of a new type of horn sound (20 times $\times$ 7 types) different from the learning data in a noisy environment. The results are shown in Tables IV and V.

As a result, by applying the integrated judgment process, we were able to obtain a classification rate of over 95% for unknown horn sounds.

D. Adding new type of data from the web

In addition to the five types of sound data collected so far (car horn, ambulance siren, bicycle bell, fire alarm, noise), we downloaded different car horns and ambulance sounds from the web page [13]. We also downloaded different bicycle bells from other web pages [14] (because the bicycle bells are not included in the [13]). We collected 428 car horn sounds, 929 ambulance siren sounds, and 169 bicycle bell sounds as new collections. Furthermore, the data was manually separated into the noisy and relatively clear data. Table VI shows the characteristics (types, numbers, and the range of sound time) of all obtained relatively clear data.

Using the CV method (5-fold, 1000 epochs) with 77 183 training and evaluation data (21 180 car horns, 28 684 ambulance sirens, 9819 bicycle bells, 12 500 fire alarms, 5 000 noises), training and evaluation were performed (input layer: 513, hidden layer: 128, output layer: 5). The classification results are shown in Table VII.

Table VII shows that the classification rates were above 94% for all NN/DNNs. However, the five-layer DNN has the highest classification rate, about 97%.

VI. LIMITATIONS OF THE DEVELOPED SYSTEM

First, we discuss data collection. The data set downloaded from the web has some problems: It includes 1) the noisy data, 2) the unlabeled data, and 3) the mixed sound data for one target (including no sound time). It is also challenging to collect significant amounts of sound data manually. This is because making a real alarm sound for acquiring such data can confuse others even when it is not truly dangerous. Crowdsourcing is a solution because crowd workers could

 ISBN: 978-1-61208-761-0

TABLE V. UNLEARNED HORN CLASSIFICATION (AFTER APPLYING THE INTEGRATED JUDGMENT PROCESS).

	TP	*FN*	*Classification rate*
Horn 1	20	0	1.00
Horn 2	20	0	1.00
Horn 3	20	0	1.00
Horn 4	20	0	1.00
Horn 5	20	0	1.00
Horn 6	20	1	0.95
Horn 7	20	1	0.95

TABLE VI. ALL DATA FOR LEARNING AND EVALUATION

	Types of alarm		*Number of each*	*Time range[sec]*
Conventional	Horn	18		6-20
Data	Bicycle bell	7	4	1-10
	Ambulance	18		1-2
Additional	Horn	151		0-4
Data	Bicycle bell	120	1-2	0-4
	Ambulance	103		0-76

record the alarm sound in daily life; other crowd workers would only label the alarm sound when they have time.

Second, in terms of the recognition response timing, a fast response time is vital because of the dangerous circumstances surrounding the sounding of alarms. There is a method to determine the recognition timing when the sound is approaching from a distance based on inverse calculation using the sound speed. However, it is difficult to distinguish the alarm sound from other environmental sounds. This problem might be solved by notifying users when the big alarm sounds occur, which happens in a hazardous situation, e.g., when the car sound is very close to the user. In this case, the way of notification is crucial.

Finally, with DHH it is difficult for people to notice the direction of the sound source. Even when the system recognizes a type of alarm sound, determining the direction of the source of that sound could be another problem. This problem could be resolved by using a microphone array and direction estimate algorithms. The mode of notifying the user of the sound direction is also essential.

VII. CONCLUSION

In this paper, we have proposed and developed an alarm sound classification system using DNNs based on smartphones. Besides, we performed evaluation experiments to verify the effectiveness of the system using the 5-fold CV, and the classification rates were above 98% for all NN/DNNs. We also proposed an integrated judgment process and made it possible to classify the types of alarms with an average F-measure of more than 99% in a real environment by using the integrated process. By applying the integrated judgment process, we were able to obtain a classification rate of over 95% for unknown horn sounds. Furthermore, even after adding the different sound data (428 car horn sounds, 929 ambulance siren sounds, and 169 bicycle bell sounds), the classification rates were above 94% for all NN/DNNs; the five-layer DNN has the highest classification rate, about 97%. We also discussed the limitations of the developed system and the expectations of the improved system by overcoming these limitations.

TABLE VII. 5-FOLD CV USING ADDITIONAL DATA.

Number of layers	*classification rate*
3	0.9367
4	0.9498
5	0.9714
6	0.9710

ACKNOWLEDGMENT

The authors would like to thank Mr. N. Hata and Mr. K. Yano, who have partially worked on the project. This work was partially supported by JSPS KAKENHI Grant Numbers #16K16460, #19K11411, and Promotional Projects for Advanced Education and Research in NTUT. One of the authors, T. Takeda, is now working at NEC Fielding, Ltd., Japan. We would like to thank Editage (www.editage.com) for English language editing.

REFERENCES

[1] World Health Organization, “Deafness and hearing loss,” March 2019, URL: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss [accessed: 2020-02-06].

[2] Fujitsu, “Ontenna,” 2020, URL: https://ontenna.jp/en/ [retrieved: February, 2020].

[3] Google Android, “live transcribe,” 2 2020, URL: https://www.android.com/accessibility/live-transcribe/ [retrieved: February, 2020].

[4] Wavio, “See sound,” 7 2019, URL: https://www.see-sound.com [retrieved: February, 2020].

[5] F. Takeda, Y. Shiraishi, and T. Sanechika, “Alarm sound classification system of oxygen concentrator by using neural network,” International Journal of Innovative Computing, Information and Control, Special Issue on Innovative Computing Methods in Management Engineering, vol. 3, no. 1, 2007, pp. 211–222.

[6] N. D. Lane, P. Georgiev, and L. Qendro, “Deepear: Robust smartphone audio sensing in unconstrained acoustic environments using deep learning,” in Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ser. UbiComp ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 283–294. [Online]. Available: https://doi.org/10.1145/2750858.2804262

[7] D. Jain, A. Lin, R. Guttman, M. Amalachandran, A. Zeng, L. Findlater, and J. Froehlich, “Exploring sound awareness in the home for people who are deaf or hard of hearing,” in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ser. CHI ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–13. [Online]. Available: https://doi.org/10.1145/3290605.3300324

[8] L. Findlater, B. Chinh, D. Jain, J. Froehlich, R. Kushalnagar, and A. C. Lin, “Deaf and hard-of-hearing individuals’ preferences for wearable and mobile sound awareness technologies,” in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ser. CHI ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–13. [Online]. Available: https://doi.org/10.1145/3290605.3300276

[9] Keras Google group., “Keras,” 9 2019, URL: https://keras.io [retrieved: February, 2020].

[10] Google, “Tensorflow,” 1 2020, URL: https://www.tensorflow.org [retrieved: February, 2020].

[11] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[12] Mitsubasankowa, “Mitsubasankowa,” 2005, URL: http://www.mskw.co.jp/car/car-horn/ [retrieved: February, 2020].

[13] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in 22nd ACM International Conference on Multimedia (ACM-MM’14), Orlando, FL, USA, Nov. 2014, pp. 1041–1044, URL:https://urbansounddataset.weebly.com/ [retrieved: February, 2020].

[14] freesound, “freesound,” 2005, URL: https://freesound.org/browse/ [retrieved: February, 2020].

 ISBN: 978-1-61208-761-0