Development of a Mobile Application for Crowdsourcing the Data Collection of Environmental Sounds

Minori Matsuyama[1], Ryuichi Nisimura[1], Hideki Kawahara[1], Junnosuke Yamada[2], and Toshio Irino[1]

[1] Wakayama University, 930 Sakaedani, Wakayama 640-8510 Japan
nisimura@sys.wakayama-u.ac.jp
[2] Nippon Telegraph and Telephone Corporation (NTT), Japan

Abstract. Our study introduces a mobile navigation system enabling a sound input interface. To realize high-performance environmental sound recognition system using Android devices, we organized a database of environmental sounds collected in our daily lives. Crowdsourcing is a useful approach for organizing a database based on collaborative works of people. We recruited trial users to test our system via a web-based crowdsourcing service provider in Japan. However, we found that improvement of the system is important for maintaining the motivation of users in order to continue the collection of sounds. We believe that the improved user interface (UI) design introduced to facilitate the annotation task. This paper describes an overview of our system, focusing on a method for utilizing the crowdsourcing approach using Android devices, and its UI design. We developed a touch panel UI for the annotation task by selecting an appropriate class of a sound source.

Keywords: environmental sound collection, user interface design, Android app, crowdsourcing.

1 Introduction

This paper introduces a mobile navigation system with a sound input interface that was developed on the basis of large-vocabulary automatic speech recognition. The system operates on Android[1] mobile devices. Figure 1 shows screenshots of our prototype system, which informs the user that there are problems in the wet area when it detects the sound of water flowing. The user can know that a patrol car is approaching when the siren is detected, as illustrated in the examples of the usage of the system depicted in Figure 2. The prototype system consists of an Android app and a web server program developed around the recognition engine.

To realize high-performance environmental sound recognition, we developed a database of environmental sounds collected from those encountered in our daily lives. A large amount of data is necessary because the recognition program

S. Yamamoto (Ed.): HIMI 2014, Part I, LNCS 8521, pp. 514–524, 2014.

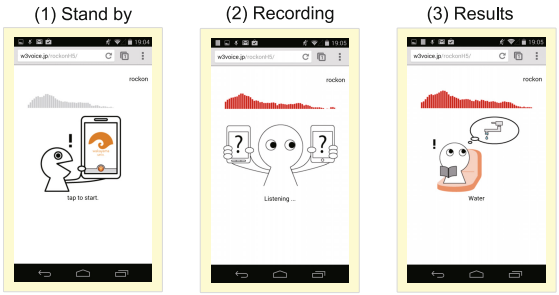

Fig. 1. Screenshots of our prototype system

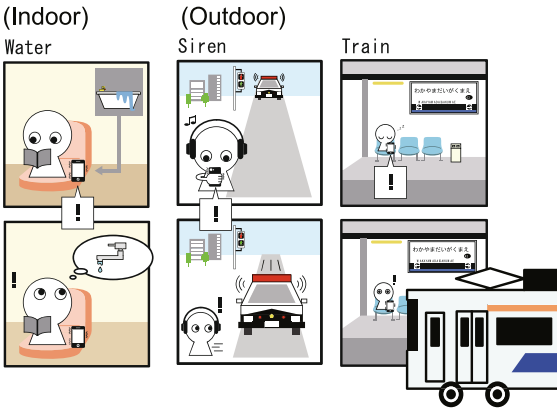

Fig. 2. Usage examples of the prototype system

utilizes a statistical pattern recognition algorithm. Our prototype system utilizes the Adaboost[2] algorithm and Hidden Markov Models (HMMs)[3] as the pattern-recognition method to identify the source of a sound. As shown in Figure 3, introduced from our previous experiments, it is necessary to improve accuracy in evaluating performances and classifying the six types of environmental sounds shown.

However,the sound data collected via Android devices in a real environment are limited. In the study reported in this paper, we succeeded in collecting sound data using a crowdsourcing approach[4]. Crowdsourcing is a practical method of employing human resources from the Internet. Further, crowdsourcing is a useful approach for developing a database on the basis of collaborative work that involves outsourcing tasks to a distributed group of people. We recruited trial users to test our system via the Rakuten research company[5], a web-based crowdsourcing service provider in Japan.

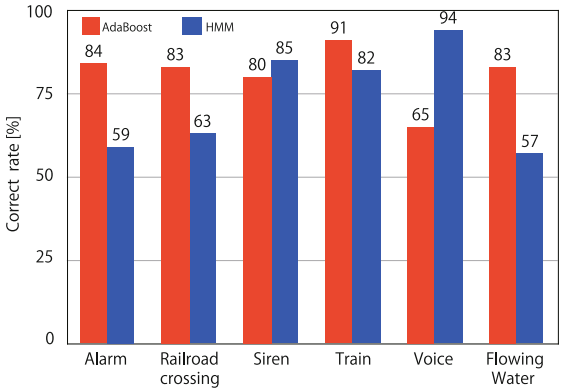

Fig. 3. Experimental results of classifying environmental sounds based on statistical pattern recognition algorithm (Our previous study)

This paper gives an overview of the environmental sound collection system, focusing on the method that utilizes the crowdsourcing approach using Android devices, and its user interface (UI) design. Data collection results are also discussed.

2 Overview of the Environmental Sound Collection System

Our system for collecting environmental sounds consists of cloud-based server-client programs. The client-side system is distributed as an Android app developed by us, which records sounds and uploads them to our web server. On the server-side, simple Common Gateway Interface (CGI)[6] programs are used to communicate with the client via HTTP[7]. The system gathers an acoustic signal for a duration of one minute, along with the GPS location information and the terminal settings with the Android OS version. Additionally, it is necessary to add metadata information, which indicates the sound-source type, to the uploaded data in order to develop a well-formed sound database. Because the system provides an annotation interface, the trial users could create the metadata information themselves.

We collected environmental sounds at the end of 2012. To collect the environmental sounds of the various regions, we first asked the Rakuten research company to prescreen trial users. We then sent the participants who registered with Rakuten research an email with the URL of the download site for our recording app. The trial users who received the email were then able to participate in the experiments by installing our app on their own Android devices. Each participant who recorded the sound and uploaded the data was given a Rakuten reward point.

Of 863 Android owners in Japan who we asked to record data, 428 sent us 841 datasets that we subsequently classified into the 92 classes shown in Figure 4.

airconditioning	cooking	jettowel	scooter	truck
airplane	cough	kettle	sea	train
alarm	crossing	lift	serverroom	TV
ambulance	crowd	lighter	shaver	vacuumcleaner
ATM	dehumidifier	mealtime	ship	vacuumtruck
baby	dog	motorcycle	sidewalk	voice
bar	doorbell	music	silent	warning
barber	drier	office	snoring	washingmachine
bicycle	dryingmachine	pachinko	snowplow	water
bird	electroniccalculator	party	stairs	waterfall
bousaimusen	elevator	patrolcar	station	WindowsOS
brass	escalator	PC	store	BAD_DATA
Buddhist altar bell	fan	piano	stove	UNKNOWN
bus	fireengine	powershovel	supermarket	
cafe	fireworks	printer	switch	
car	game	radio	telephone	
cat	gamecenter	railwaycrossing	temple	
chime	horn	rain	ticketmachine	
clippers	horse	river	toilet	
clock	hospital	runner	toy	

#>=10 #>=20 #>=30 #>=40 #>=50

#>=60 #>=70 #>=80

Fig. 4. Table of 92 classes defined by us, where the color of each cell indicates the number (#) of datasets collected

The overall recording time of the uploaded data was 10 hours five minutes. The classes were determined according to the type of the sound source (e.g., train, car, or bird).

Among the datasets obtained, we were not able to record the GPS position information for 253 because at the time those datasets were being uploaded, the data got corrupted as a result of network problems. In addition, because of problems associated with the accuracy of the GPS sensor of user's Android terminal, we confirmed that the GPS position information could not be acquired before the start of any recording.

3 UI Design for the Recording App on Android Mobile Devices

Crowdsourcing proved useful for easily developing the sound database. However,we discovered that improvements to the system were necessary to maintain the motivation of trial users in order for them to continue the sound collection activity. We believe that the enhanced UI introduced to facilitate the annotation task addressed this problem. Figure 5 shows our original UI for the annotation task; a text-input form is displayed for the user to input the sound source type. However, because some software keyboards on Android devices are too small, users often feel that it is too difficult to annotate the data.

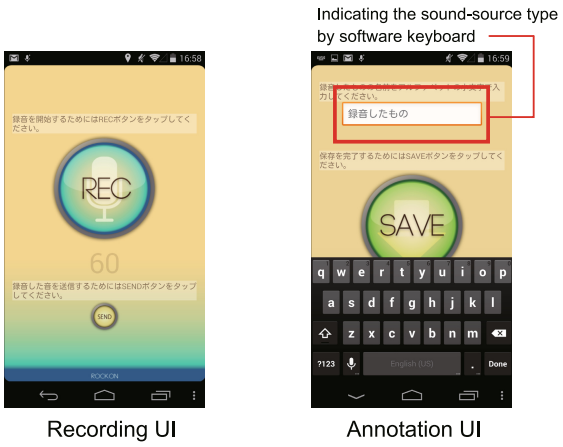

Fig. 5. Original UI for the annotation task to decide on a sound source type via the software keyboard

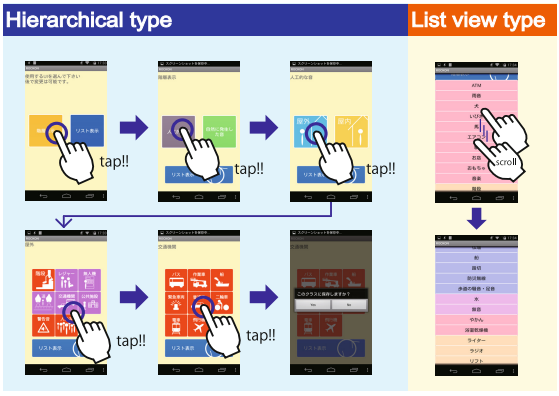

Fig. 6. Screenshots of new UI for selecting the appropriate sound-source class from the candidates

To overcome this problem, as demonstrated in Figure 6, we developed a new UI that enables users to simply select an appropriate sound source class from a list prepared in advance. In the new sound source annotation interface, we provided two types of UIs: a hierarchical type and a list view type. Table 1 compares the two types of UIs. When using the hierarchical type UI, the user can choose a class via touch panel operations by tracing a hierarchical tree of sound-source classes. We believe that the system can support step-by-step operations to choose the sound source, although the user does not know the entire class structure. In the list view, the candidate sound sources are arranged, and each user can select a sound source by scrolling the candidate list to the appropriate choice.

Table 1. Our basic design of the sound source annotation UI

Hierarchical type
Hierarchical tree view of the sound source classes structured in groups. The names of the groups with the higher level of abstraction are applied to ensure diversity of environmental sounds during recordings. To reduce the complexity of the operation, the depth of the hierarchy is limited to four levels.
List view type
A long list is presented without the candidates of the sound source class being separated. Users can select a sound source by scrolling the list displayed in alphabetical order. The abstracted groups are not directly introduced for sound-source class candidate selection.

Table 2. List of sound sources used in the annotation experiment

Large-number classes	Small-number classes
train, car, railway crossing, TV, voice, station, crossing, store, bird, rain.	cat, electronic calculator, piano, river, sea, ship, warning, lighter, stairs, telephone.

To evaluate our new UI, we developed a prototype application that runs on Android smartphones. Ten participants attempted to choose the appropriate class of a sound source after they listened to environmental sounds under the condition that the participants did not know the situations in which the sounds were recorded. The test samples for each participant were the 20 sounds listed in Table 2. Among them, 10 samples were extracted from the sound-source classes with a large number of sounds recorded, and 10 samples belonged to the classes that had only a small number of sounds recorded.

We counted the number of steps (the number of screen taps) and the elapsed time required to determine the class after listening to the audio signal. Tables 3 and 4 show the elapsed times and the number of steps, respectively.

From these results, it can be seen that the times required for decisions using the list view type are longer than those for the hierarchical type. We believe that the participants were able to imagine the source of the sounds by listening to the acoustic signals. When looking for a candidate from the list, the participants tended to carefully choose the class over time. On the other hand, in the hierarchical type UI, it is possible that the time required to browse classes was shorter because only a limited number of candidates was displayed on one screen.

Because in the list view type, a class can only be selected via the scroll operation, the number of steps is shown as approximately one. In the hierarchical type, the number of steps for the 10 classes with the highest number of samples is less than that for the small-number classes. For the 10 classes with the smallest number of samples, we observed that the participants could not identify the origin of the sound source while looking for a suitable class.

Table 3. Elapsed time [s] (Average)

	Large-number class	Small-number class
List view type	17.49	19.50
Hierarchical type	13.64	18.46

Table 4. Number of steps (Average)

	Large-number class	Small-number class
List view type	1.13	1.06
Hierarchical type	5.04	6.40

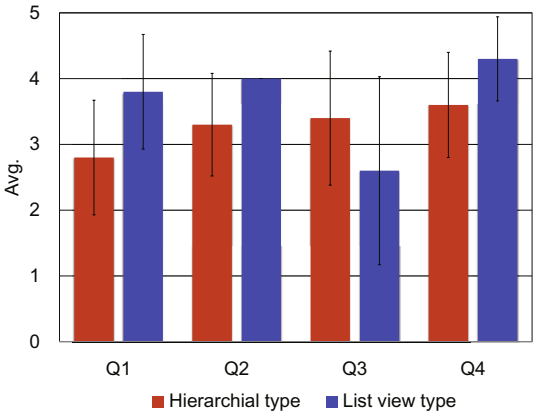

Fig. 7. Results of the five-grade evaluations of UI design

Next, we conducted five-grade evaluations in which we interviewed the participants. The following questions were posed to the participants:

Q1: Did you hesitate in selecting a class? (1: many times – 5: no)
Q2: Did the selected class match your image? (1: no match at all – 5: almost matched)
Q3: Were your operations affected by the visibility of the UI? (1: poor visibility, difficult – 5: good visibility, easy operation)
Q4: Could you operate using only the preliminary instructions? (1: very difficult – 5:very easy)
Q5: Which UI design did you prefer? (”Hierarchical type” or ”List view type”)

Evaluation results for Q1 to Q4 are shown in Figure 7. The bars display the averaged values from 10 participants, where the left bars are the results for the hierarchical type UI, and the right bars are for the list view type UI. Standard deviations are also indicated in the error bars.

In Q1 and Q2, the list view type UI received a high evaluation. Because the sound source classes presented in the hierarchical type are abstracted, the user may not be able to conceive of the lower layer classes. Using the list view type UI, however, the entire list of sound sources can be viewed. Therefore, the system did not give room to allow guessing so the participants hesitated to select one. We obtained high evaluations for the hierarchical type UI via Q3. The results demonstrate that the hierarchical type UI that reduces the amount of information simultaneously displayed on one screen is easy to see. We believe that the reason for the high evaluation of list view type in Q4 is the fact that the list view type UI has a familiar design that is similar to that of the smartphone apps the participants use on a daily basis.

In Q5, six participants chose the list view type UI, whereas four preferred the hierarchical type UI. Consequently, we concluded that there is no significant difference between both UIs in terms of convenience. In order to utilize the advantages of both types, we implemented an annotation UI that can be switched between both types of UIs.

4 Experimental Collection of Sounds in Real Environments

This section describes a small-scale environmental sound collection experiment performed preliminarily by using a modified application, which includes an annotation UI that can be switched between the list view type and the hierarchical type UIs. The experiment conducted asked that each participant select the sound source class without knowing the recording status of the environmental sound in Section 3. In addition to the information from the ear, visual information from the eye gives the impression of change in environmental sounds. In this study, we conducted trial tests of the improved application in the real world in order to investigate the affect of visual information. However, this experiment was small scale because it was only a preliminary investigation. Large-scale experiments with crowdsourcing will be conducted in the future.

The participants in the experiment were 11 students who are Android device users. The experiment was conducted over a period of three days, starting February 3, 2014. We asked the participants to record the environmental sounds they encountered in everyday life. However, we did not set a quota on the number of collections per individual. As a result, we were able to collect 79 environmental sound datasets.

In addition, we conducted interviews with the participants after this experiment. We asked the following questions in the interview:

Q1: Were your operations affected by the visibility of the UI? (1: poor visibility, difficult – 5: good visibility, easy operation)
Q2: Were you able to easily operate the app? (1: very difficult – 5: very easy)
Q3: Did you hesitate in selecting a class? (1: many times – 5: no)
Q4: Did the selected class match your image? (1: no match at all – 5: almost matched)

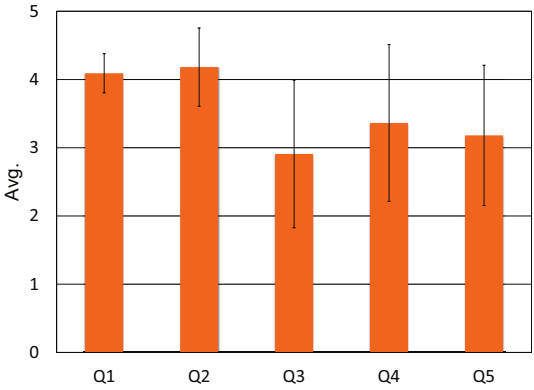

Fig. 8. Results of the five-grade evaluations in the environmental sound collection experiment

Q5: Do you think that the candidate sound sources displayed were appropriate? (1: not appropriate – 5: appropriate)

The results obtained are depicted in Figure 8, where the evaluation values given from the 11 participants are averaged. We obtained evaluation values greater than four for the items visibility (Q1) and operability (Q2). The validity of the improved UI was also confirmed. The evaluation value for both Q3 and Q4 was approximately three. This is a result of a tendency that is different from the similar entry in Section 3. The cause of this difference lies in the fact that the participants were finding the target sound source that they would record as environmental sounds by themselves. Because the participants in Section 3 heard environmental sounds collected in advance, they could not have any material for determining the status of the peripheral as visual information. Therefore, the annotation task was actually one of guessing the sound source while viewing the candidates presented on the screen. In contrast, this sound collection experiment required that the participants have a concrete image of the sound source in order to perform the annotation task to determine the appropriate sound source class. We believe that if the sound source the participants expected could not be found on the screen, it became a worrying factor in determining the class.

In addition, we investigated the impression participants had of the collection tasks on the basis of a single representative word from each person. We presented the 14 words listed in Table 5 to express the impression (impression words) for them, and asked that they each select the word closest to the impression they felt. The set of impression words comprise seven positive words and seven negative words. Multiple answers for the impression words were allowed.

Figure 9 shows the number of times each impression word was selected. As can be seen in the figure, in many cases, participants had positive impressions associated with the use of our app. These results play an important role because keeping participants motivated is essential when crowdsourcing is used.

Table 5. 14 impression words in Japanese and English (translated)

Positive	楽しい (enjoyable)	心地よい (comfortable)	おもしろい (amusing)	めずらしい (unusual)
	新しい (novel)	賢い (smart)	さりげない (casual)	
Negative	うっとうしい (depressing)	難しい (difficult)	いいかげん (irresponsible)	なぞ (ambiguous)
	つらい (miserable)	恥ずかしい (ashamed)	めんどう (troublesome)	

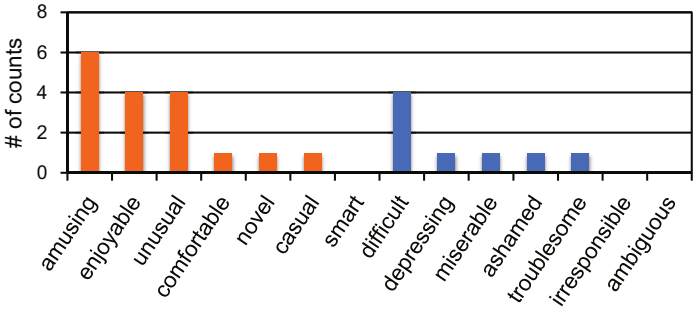

Fig. 9. Results for impression word evaluation in the environmental sound collection experiment

However, four participants had a "difficult" impression; therefore, further improvements to the system are necessary.

5 Conclusions and Future Work

This paper described an overview of our environmental collection system focusing on a method for utilizing the crowdsourcing approach using Android devices. We developed a touch panel UI for the annotation task by selecting an appropriate class of a sound source. The annotation system is composed of two types of UIs: a hierarchical type and a list view type.

We are planning to perform field tests of our improved sound collection system based on the crowdsourcing approach. We hope to involve many participants that are accustomed to working on advanced UIs to conduct the annotation work efficiently in our experiments. To realize this, we believe it is necessary to consider introducing shortcut and search functions to the sound source annotation UI.

Acknowledgments. This study was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (KAKENHI), Japan. We greatly appreciate the support provided by Mr. Takahiko Tsuda and Mr. Kyosuke Nakanishi (graduates of our laboratory).

References

1. http://www.android.com/
2. Freund, Y., Schapire, R.E.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)
3. Lee, A., et al.: Noise Robust Real World Spoken Dialogue System using GMM Based Rejection of Unintended Inputs. In: Proc. of INTERSPEECH, pp. 173–176 (2004)
4. Parent, G., Eskenazi, M.: Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges. In: Proc. of INTERSPEECH, pp. 3037–3040 (2011)
5. http://research.rakuten.co.jp/en/
6. Robinson, D., et al.: The Common Gateway Interface (CGI) Version 1.1. RFC 3875. IETF (Internet Engineering Task Force) (2004)
7. Fielding, R., et al.: Hypertext Transfer Protocol - - HTTP/1.1, RFC 2616. IETF (Internet Engineering Task Force) (1999)