Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

The Acoustic Sound Field Dictation with Hidden Markov Model Based on an Onomatopeia

Kohei Hayashida (1), Yu Mizoguchi (1), Junpei Ogawa (2), Masanori Morise (2), Takanobu Nishiura (2), and Yoichi Yamashita (2)

(1) Graduate School of Science and Engineering, Ritsumeikan University, Kusatsu, Japan
(2) College of Information and Science, Ritsumeikan University, Kusatsu, Japan

PACS: 43.72.NE

ABSTRACT

In this study, we realized acoustic sound field dictation which is effective for the security systems because it can quickly find an abnormal sound on the basis of text information from a captured long signal. Environmental sound identification has been previously researched only with the method that individually models all sound sources. However, it is impossible to model the innumerable real world environmental sounds. Therefore in our research, we try to reduce the number of models by utilizing onomatopoeias because they can represent an acoustic sound as a word. We thus firstly aimed at the environmental sound identification with the hidden Markov model (HMM) on the basis of onomatopoeias for realizing acoustic sound dictation. We carried out the subjective evaluation experiment with identification results of real world acoustic sounds. The results confirmed that the proposed method can better use identification result to remind people of acoustic sound than the conventional method.

INTRODUCTION

Advances in computer technology have improved an acoustic sound fields understanding [1]. Acoustic sound field understanding enables the acoustic sound field dictation, the automatic acoustic sound identification, and the automatic acoustic sound recognition with higher performance. In addition, it is possible to archive the acoustic sound field with high quality if the acoustic sound field dictation is realized that includes not only the human voice but also the environmental sounds. Thus in this study, we created an acoustic sound field dictation system. This system is effective for the security systems because it can quickly search for an abnormal sound on the basis of text information from a captured long signal. It is indispensable for realizing the acoustic sound field dictation system. Environmental sound identification has been previously researched only with the method that individually models all sound sources [2]. However, it is impossible to model the innumerable real world environmental sounds. Therefore in our research, we reduce the number of the models by utilizing onomatopoeias because they can represent an acoustic sound as a word. We thus firstly aimed to identify environmental sounds with the hidden Markov model (HMM) [3] on the basis of onomatopoeias for realizing acoustic sound dictation.

CONVENTIONAL METHOD FOR ENVIRONMENTAL SOUND IDENTIFICATION

We are surrounded by various environmental sounds: passing trains and cars, chirping birds or insects, ringing cellular phones, and so on. The method based on the hidden Markov model (HMM) [3] has been developed to identify these environmental sounds [2]. Environmental sound identification has only been researched with the method that individually models all sound sources. However, it is impossible to model the innumerable environmental sounds in the real world. In addition, identifying similar sounds using the same category is useful for searching in the sound field dictation system. In our research, we developed the HMM on the basis of onomatopoeias to reduce the number of models.

ENVIRONMENTAL SOUND IDENTIFICATION WITH HMM BASED ON AN ONOMATOPEIA FOR PROPOSED METHOD

The words that imitate various natural sounds are defined as onomatopoeias [4]. Recently, the relationships among the sounds, the onomatopoeias and human senses have been studied [5 ~ 7]. These studies show that the onomatopoeia can remind someone of an acoustic sound. Therefore, we developed an identification method for environmental sounds in which similar sounds are categorized in the same category on the basis of onomatopoeias. Figure 1 shows the basic concept of the conventional and proposed methods. As shown in Fig. 1, the proposed method can represent similar sounds in one category by using an onomatopoeia model. In this study, we aim to reduce the number of models and realize an easy-to-understand acoustic sound field dictation by our proposed method.

PRELIMINARY EXPERIMENTS OF THE OPTIMUM PARAMETERS AND THE ONOMATOPOEIC CATEGORIES

In this study, we carried out two preliminary experiments: one to find out optimum parameters for the identifying environmental sounds and the other to find out the

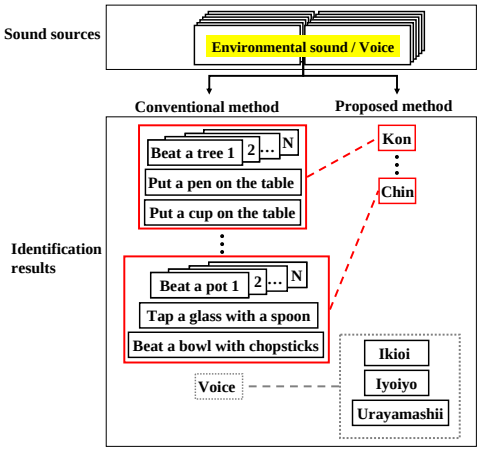

Figure 1. Basic concept of conventional and proposed methods

correspondence relationships between the environmental sounds and the onomatopoeias.

The Examination of the Optimum Parameters for Identification

First, we investigated the optimum parameters for environmental sounds identification. We used the sound scene database in a real acoustic environment that is recorded by real world computing partnership (RWCP-DB) [8] to train the models and evaluate the proposed method.

Optimum Parameter of Sampling Frequency

Figure 2 shows the relationship between the identification rate, mel-Frequency cepstrum coefficient (MFCC) order, and sampling frequency in 5 states and 128 mixtures based on HMM. Figure 3 shows the relationship between identification rate, MFCC order, and sampling frequency in 8 states and 128 mixtures. Figure 4 shows the relationship between the identification rate, MFCC order, and sampling frequency in 10 states and 128 mixtures. Figures 2 ~ 4 show that the high identification accuracy is in an 8 ~ 16 [kHz] sampling frequency. Figure 5 shows the identification rate without that of 10 states, because the identification accuracy in 10 states is lower than others. As the results in Fig. 5 confirm that the optimum parameter of sampling frequency is 16 [kHz].

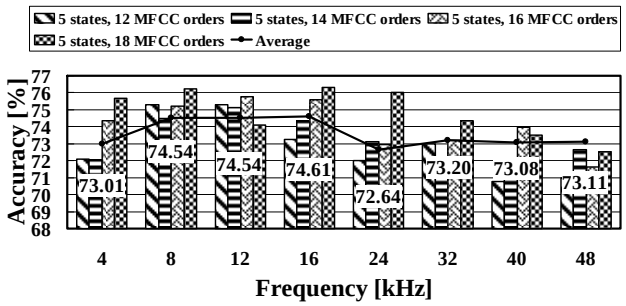

Figure 2. Relationship between identification rate, MFCC order, and sampling frequency in 5 states.

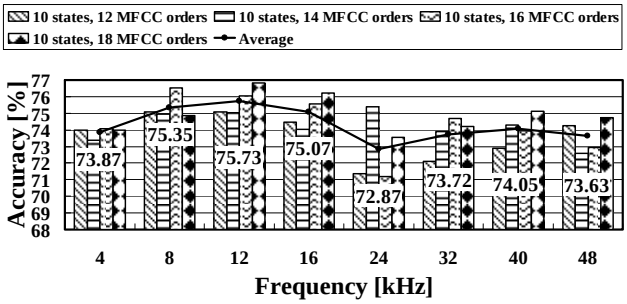

Figure 3. Relationship between identification rate, MFCC order, and sampling frequency in 8 states.

Optimum Parameter of MFCC Order

We used 16 [kHz] sampling frequency confirmed in the previous section. Figure 6 shows the relationship between the identification rate and MFCC order in 5 states and 128 mixtures. Figure 7 shows the relationship between the identification rate and MFCC order in 8 states and 128 mixtures. Figure 8 shows the relationship between the identification rate and MFCC order in 10 states and 128 mixtures. Figure 9 shows the relationship between the identification rate and MFCC order in 128 mixtures. The results confirmed that the optimum parameter of MFCC order is 16.

The Optimum Parameters of State and Mixture

We used 16 [kHz] sampling frequency and 16 MFCC orders. Figures 10 and 11 show the relationship between the identification rate and numbers of states and mixtures. The results in Fig. 10 confirm that the optimum parameter of the state is 8. The results Fig. 11 confirm that the optimum parameter of the mixture is 128.

As the result, we confirmed that the optimum parameters for HMM based on the onomatopoeia are a 16 [kHz] sampling frequency, 16 orders MFCC, 8 states, and 128 mixtures.

The Examination of the Onomatopoeia Category for Identification

Next, we investigated the correspondence relationship between the environmental sounds and the onomatopoeias using questionnaires. We classified 92 kinds of sound source

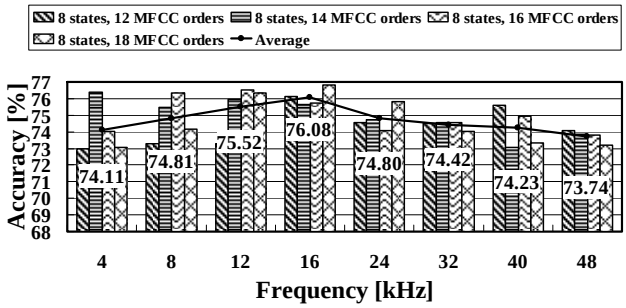

Figure 4. Relationship between identification rate, MFCC order, and sampling frequency in 10 states.

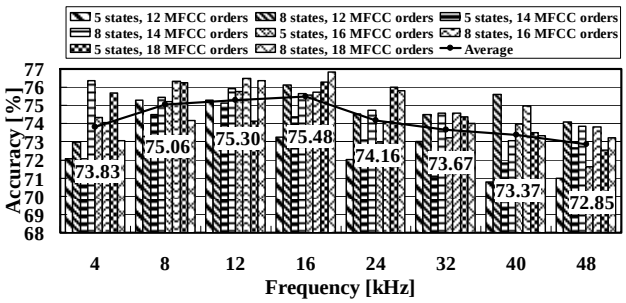

Figure 5. Relationship between identification rate, MFCC order, and sampling frequency in 128 mixtures.

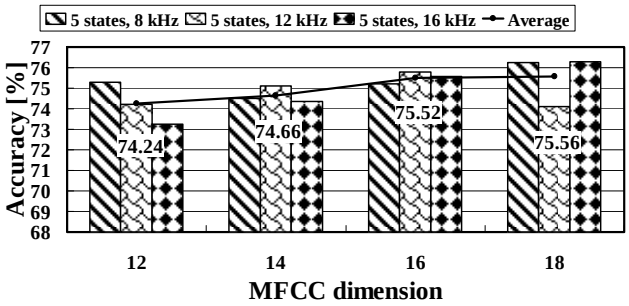

Figure 6. Relationship between identification rate and MFCC order in 5 states and 128 mixtures.

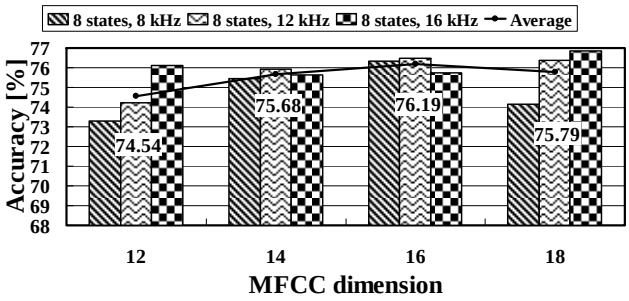

Figure 7. Relationship between identification rate and MFCC order in 8 states and 128 mixtures.

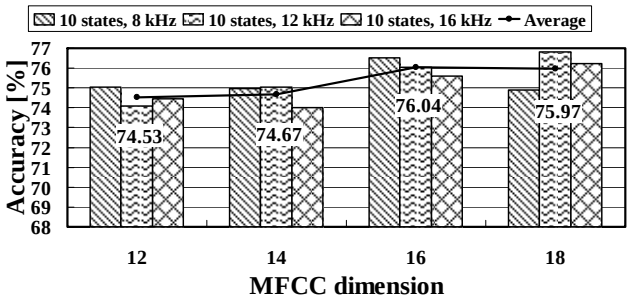

Figure 8. Relationship between identification rate and MFCC order in 10 states and 128 mixtures.

in RWCP-DB into the onomatopoeia categories. The subjects were 2 females and 15 males. The subjects had 12 choices in each kind of sound and chose a suitable onomatopoeia to express a presented sound. We presented the onomatopoeia choices for each sound to subjects on the basis of the Hiyane's classification of RWCP-DB [9]. Figure 12 shows questionnaire format. Using the results of the questionnaire, the environmental sounds of RWCP-DB are classified into 33 onomatopoeia categories shown in Table 1.

EVALUATION EXPERIMENTS

In this study, we carried out two evaluation experiments: one is the identification experiment of the environmental sounds, and the other is the subjective evaluation experiment using identification results of real world acoustic sound.

Identification Experiment of the Environmental Sounds

Experimental conditions

We compared the identification accuracy of the conventional method with that of the proposed method. The conventional method identifies the environmental sound source on the basis of the RWCP standard category (Tab. 2). The proposed method identifies the environmental sound source based on the onomatopoeia category (Tab. 1). Table 3 shows the

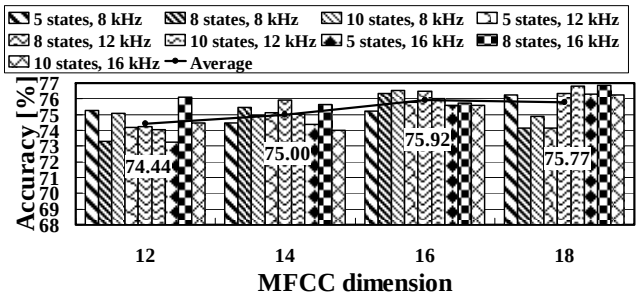

Figure 9. Relationship between identification rate and MFCC order in 128 mixtures.

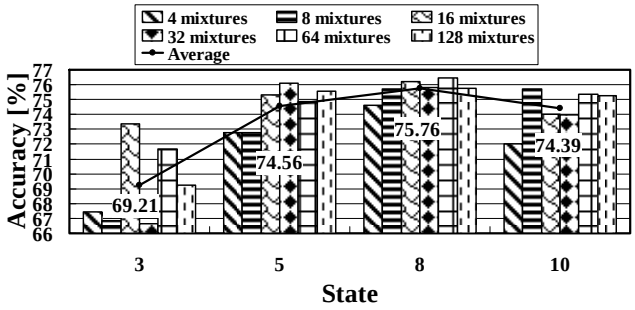

Figure 10: Relationship between identification rate and state.

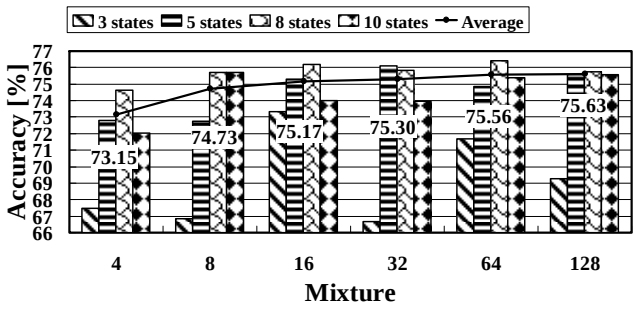

Figure 11: Relationship between identification rate and mixture.

	Kochi	Kon	Tan	Ton	Doshi	Don	Bon	Cha	Chi	Kata	Koto	Pata
cherry1												
cherry2												

Figure 12: Questionnaire format for preliminary experiment.

Table 1. Onomatopoeia category

Bashi	Bu:	Cha	Charin	Chi	Chirin
Gan	Gasa	Go:	Kacha	Kan	Karan
Kanra	Kata	Kata-kata	Katata	Kyu	Pa:n
Paki	Pan	Pi:	Pinpon	Pipipi	Piriri
Piroro	Po:n	Shu	Shu:	Chin	Ton
Za	Za:	Jiriri			

experimental conditions. We used 86 sources in RWCP-DB without insufficient sound of number for the training of HMM and the environmental sound identification.

Experimental Results

We compared the identification error rate of the conventional method with that of the proposed method. Figure 13 shows the identification error rate of the conventional method (RWCP-DB standard category), and Fig. 14 that of the proposed method (Onomatopoeia category). According to the results in Figs. 13 and 14, the proposed method has a lower identification error rate than the conventional method. Therefore, we confirmed the effectiveness of the proposed

Table 2. RWCP standard category

cherry1	cherry2	cherry3	mang1	mang2	mango3
teak1	teak2	teak3	wood1	wood2	wood3
bank	bowl	candy bw1	coffcan	colacan	metal05
metal10	metal15	pan	trashbox	case1	case3
dice1	dice2	dice3	bottle1	bottle2	china1
china2	china3	china4	cup1	cup2	particl1
particl2	pump	spray	file	sandpp1	sandpp 2
saw1	saw2	aircap	sticks	cap1	cap2
clap1	clap2	claps1	claps2	snap	bells5
coin1	coin2	coin3	coins1	coins4	book1
book2	crumple	tear	castanet	drum	horn
maracas	ring	string	whisle1	whisle2	whis-tle3
buzzer	clock2	phone1	phone4	pipong	clock1
coffmill	doorlock	dryer	mechbell	padlock	punch
shver	stapler				

Table 3. Experimental conditions environmental sounds identification.

Sampling frequency	16 kHz
Feature vector	33 orders 16 orders MFCC + 16 orders ΔMFCC + 1 order ΔPower
Number of state	8 states
Number of mixture	128 mixtures
Data for traning	86 kinds of sounds in RWCP-DB, 50 samples
Data for Identification	86 kinds of sounds in RWCP-DB, 50 samples

method. The proposed method tends to misconstrue "Shu " as "Za " and "Karan" as "Kan". The results of the subjective evaluation denoted the same tendency as the questionnaire. Therefore, the sound sources that are difficult for people to identify are also difficult for a computer to identify.

Subjective Evaluation with Identification Result

Experimental Conditions

Table 4 shows the recording conditions. Twenty kinds of sound source for the subjective evaluation were recorded by five times each. Table 5 shows the sound source number and the type of sound source. After the identification of real world environmental sounds by the conventional and proposed methods, we investigated whether onomatopoeia can remind someone of the sound source from the identification results. We carried out the subjective evaluation with three choices: can remind, cannot remind, or neither. We used the identification results for subjective evaluation in which the identification accuracy was over 60 % by both methods. The subjects were two females and 13 males. Figure 15 shows the questionnaire format for the subjective evaluation.

Experimental results

Figures 16 and 17 show average rate and standard deviation of the opinions in the subjective evaluation for the conventional and proposed methods. The results confirmed that the proposed method can better use identification result to remind people of acoustic sound than the conventional method. Therefore, the proposed method accurately realized the easy-to-understand acoustic sound field dictation.

However, more people were not reminded of an acoustic sound than those who were in 7: closing a drawer, 10: closing a sliding door, 15: closing a door, and 16: closing a window (Tab. 6). The sounds of 7, 10, 15, and 16 are composed of

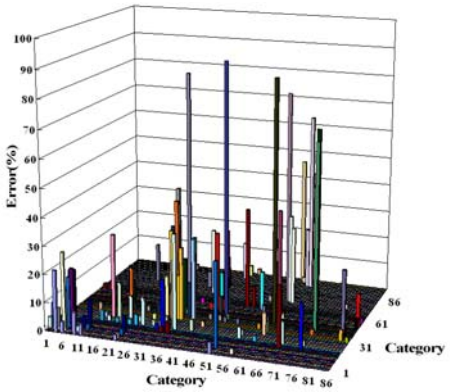

Figure 13: Identification error rate of the conventional method (RWCP standard category).

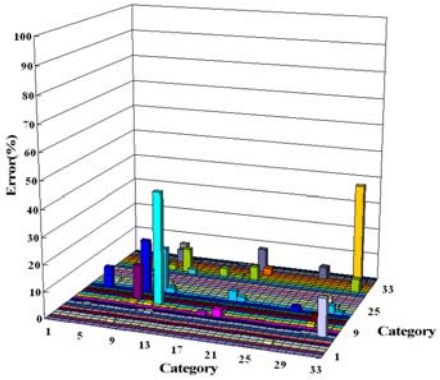

Figure 14: Identification error rate of proposed method (Onomatopoeia category).

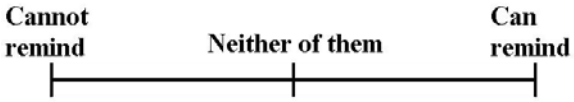

Figure 15: Questionnaire format for subjective evaluation.

two syllabics. Figure 18 shows the time waveform of sound 7: close a drawer. As the power of these sound sources concentrates on the two times, these sound sources are expressed by the onomatopoeia on the basis of two syllabics. Therefore, the proposed method of utilizing a syllabic onomatopoeia is insufficient for training models in the complex sounds in real environments. In addition, we need to consider the production method for models of the complex sounds by utilizing onomatopoeia on the basis of over two syllabics.

CONCLUSIONS

In this study, we proposed a hidden Markov model (HMM) based on the onomatopoeia for environmental sound identification. As the result of the experiment of environmental sound identification, we confirmed that the proposed method can more accurately identify sounds than the conventional method. As a result of the subjective evaluation experiment, we confirmed that the proposed method can easily remind an acoustic sound with an identification result compared with the conventional method. Therefore, the proposed method accurately realized an easy- to-understand acoustic sound field dictation. However, we confirmed that the proposed method cannot build sufficient models of complex environmental sounds. In future work, we will try to build such models. Therefore, we need to record a lot of complex environmental sounds and develope a database for them.

ACKNOWLEDGEMENTS

This work was partly supported by Global-COE and Grants-in-Aid for Scientific Research funded by The Japanese Ministry of Education, Culture, Sports, Science and Technology.

REFERENCES

1 Hiroshi G. Okuno, "Auditory scene analysis from the viewpoint of computer science," *The Journal of Acoustical Society of Japan Vol. 50, no. 12*, pp. 1017—1022 1994.
2 Kazuhiro Miki, *et. al.*, "Environmental sound discrimination based on hidden markov model," *Technical Report of IEICE. SP99-106*, pp. 79—84, 1999.
3 Keiichi Tokuda, "Fundamentals of speech synthesis based on HMM," *Technical Report of IEICE. SP2000-74*, pp. 79—84, 2000.
4 Ikuhiro Tamori, "Characteristics of Japanese onomatopoeia," *The Journal of Acoustical Society of Japan, Vol. 54, no. 3*, pp. 215—222, 1998.
5 Keiji Kawai, *et. al.*, "Personal psychological evaluation structure of environmental sounds: Experiments of subjective evaluation using subjects' own terms," *The Journal of Acoustical Society of Japan, Vol. 60, no. 5*, pp. 249—257, 2004.
6 Yuichi Kato, *et. al.*, "A proposal of the system for rating timbre based on onomatopoeia," *Proc. of Acoustical Society of Japan*, pp. 725—726, Sep. 2003.
7 Nozomu Fjisawa, *et. al.*, "Study on auditory imagery associated with onomatopoeic representation," *Technical Report of IEICE., SP2003-196*, pp. 19—24, 2004.
8 Satoshi Nakamura, *et. al.*, "Design and status of sound scene database in real acoustical environments," *Proc. of Acoustical Society of Japan*, pp. 137—138, Sep. 1998.
9 Kazuhiro Hiyane, *et. al.*, "Study of spectrum structure of short-time sounds and its onomatopoeia expression,", *Technical Report of IEICE. SP97-125*, pp. 65—72, 1998.

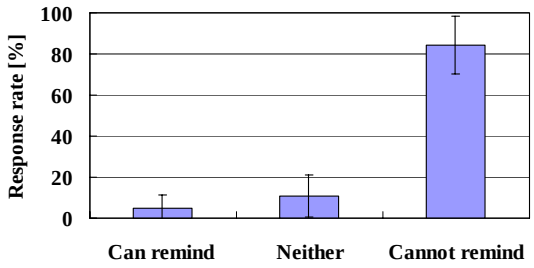

Figure 16: Results of subjective evaluation of conventional method.

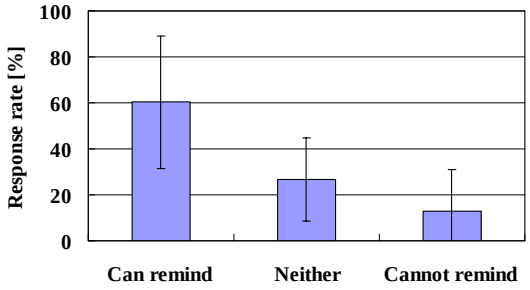

Figure 17: Results of subjective evaluation of proposed method.

Table 4. Recording conditions for real world environmental sounds.

Ambient noise level	47.0 dBA
Temperature	26 deg C
Humidity	8.0 %
Recorder	SONY, PCM-D1

Table 5. List of recorded sounds in real environment.

Sound number	Type of sound source
1	Ringtone of a cellphone
2	Putting a glass on the table
3	Tapping a glass with a spoon
4	Rattling a spoon in a glass
5	Putting a can on the table
6	Dropping a spoon on the floor
7	Closing a drawer
8	Picking up a key with the bell
9	Stapling paper together
10	Closing a sliding door
11	Dropping a pen on the floor
12	Pressing the enter key
13	Footsteps
14	Sitting down on a chair
15	Closing a door
16	Closing a window
17	Cell phone on silent mode
18	Crumpling paper into a ball
19	Tearing paper
20	Spray can sound

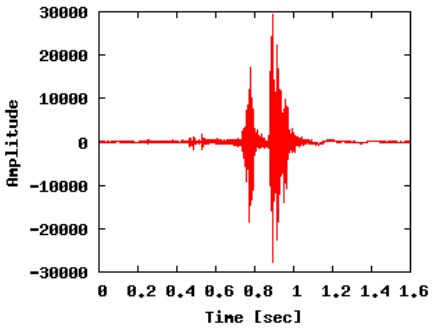

Figure 18: Time waveform of sound 7: "closing a drawer".

Table 6. Results of identification and subjective evaluation (proposed method).

		Subjective evaluation result		
Sound number	Ident. result	Can remind	Neither	Cannot remind
1	Pirii	15	0	0
2	Gan	12	3	0
3	Kan	10	5	0
4	Kan	8	7	0
5	Kan	15	0	0
6	Karara	13	1	1
7	Gan	2	4	9
8	Karara	7	8	0
9	ZA	10	3	2
10	Gan	5	5	5
11	Gan	12	2	1
12	Paki	11	3	1
13	Gan	13	2	0
14	ZA	12	3	0
15	Gan	2	7	6
16	Gan	2	10	3
17	Zai	9	5	1
18	Gasa	13	2	0
19	Gasa	5	3	7
20	Gasa	5	7	3