

ENVIRONMENTAL SOUND SOURCE IDENTIFICATION BASED ON HIDDEN MARKOV MODEL FOR ROBUST SPEECH RECOGNITION

Takanobu Nishiura^{†,‡}, Satoshi Nakamura[†], Kazuhiro Miki[§] and Kiyohiro Shikano[§]

[†] ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Keihanna Science City Kyoto, 619-0288 Japan

[‡] Faculty of Systems Engineering, Wakayama University

[§] Graduate School of Information Science, Nara Institute of Science and Technology
{takanobu.nishiura@atr.co.jp}

ABSTRACT

In real acoustic environments, humans communicate with each other through speech by focusing on the target speech among environmental sounds. We can easily identify the target sound from other environmental sounds. For hands-free speech recognition, the identification of the target speech from environmental sounds is imperative. This mechanism may also be important for a self-moving robot to sense the acoustic environments and communicate with humans. Therefore, this paper first proposes Hidden Markov Model (HMM)-based environmental sound source identification. Environmental sounds are modeled by three states of HMMs and evaluated using 92 kinds of environmental sounds. The identification accuracy was 95.4%. This paper also proposes a new HMM composition method that composes speech HMMs and an HMM of categorized environmental sounds for robust environmental sound-added speech recognition. As a result of the evaluation experiments, we confirmed that the proposed HMM composition outperforms the conventional HMM composition with speech HMMs and a noise (environmental sound) HMM trained using noise periods prior to the target speech in a captured signal.

1. INTRODUCTION

It is very important for the natural interfaces of machines such as self-moving robots to capture distant-talking speech with high quality and to recognize such speech accurately. However, environmental sound noise, background noise and room reverberations seriously degrade the sound capture quality in real acoustical environments. In turn, automatic speech recognition (ASR) performance degrades in such environments. To overcome this problem, in this paper, we focus on environmental sound source identification toward robust speech recognition in noisy environments.

For a robust speech capture technique, a microphone array is an ideal candidate as an effective method of capturing distant-talking speech. The desired speech signals can be selectively acquired by precisely steering the microphone array in the desired speech direction [1]. However, this approach requires localizing the target talker. Conventional talker localization algorithms in multiple sound source environments not only have difficulty localizing the multiple sound sources accurately, but also have difficulty localizing the target talker among localized multiple sound source positions. In this case, if environmental sound noise can be identified, talker localization is expected to be more accurate.

For a robust speech recognition technique, many approaches

have been proposed. A simple and powerful technique is the Spectrum Subtraction (SS) [2], which was proposed by S. F. Boll in 1979. The SS technique reduces the noise signal by subtracting the noise spectrum from the noisy speech spectrum after estimating the noise spectrum. However, although it has a good subtracting performance in a stationary noisy environment, it is not sufficient in a non-stationary noisy environment. On the other hand, the Hidden Markov Model (HMM) composition [3, 4] method has also been proposed as a method to improve ASR performance. This method adapts an HMM to the quasi-noisy-environment before recognizing the noisy speech. Therefore, it is very important for selecting the quasi-environment to estimate the kind of environmental sound noise in the real environment while composing an HMM to a quasi-noisy-environment. In this case, if the environmental sound noise can be identified, the HMM composition is more effective for robust speech recognition.

In this paper, we focus on the environmental sound source identification toward robust speech recognition and study the environmental sound source modeling with HMMs.

2. ENVIRONMENTAL SOUND SOURCE DATABASE

Numerous environmental sound sources are necessary for this research. Therefore, we employ the RWCP (Real World Computing Partnership) sound scene database (RWCP-DB) [5]. The RWCP-DB includes a lot of environmental sound sources that are collected in an anechoic room. Table 1 shows the types of environmental sound sources. The first category is collision sounds of wood, plastic and ceramics. The second and the third category are composed of sounds that occur when humans perform to contain activities, such as, spraying, sawing, clapping, and the sounds generated when using coins, books, pipes, telephones, toys, and so on. The sounds of the second category are the sounds whose sound source materials can not easily be associated. Whereas, the source materials of the third category sounds can be easily associated uniquely. Approximately 100 samples for about 92 kinds of sounds sufficient enough for statistical model training are recorded.

3. ENVIRONMENTAL SOUND SOURCE IDENTIFICATION

We carried out evaluation experiments of environmental sound source identification with the environmental sound HMM.

Table 1: Environmental sound sources (Dry source sounds)

Category	#Samp.	Sound source
Collision Sound		
Wood	1187	wood boards, wood stick
Metal	1000	metal boards, metal stick
Plastic	550	plastic boards, plastic stick
Ceramic	800	glasses, china
Action Sound		
Article dropping	200	dropping article in box
Gas jetting	200	spray, pump
Rubbing	500	sawing, sanding
Bursting and breaking	200	breaking stick, air cap
Clapping sound	829	hand clap, slamming clip
Characteristic Sound		
Small metal articles	1072	small bell, coin
Paper	400	dropping book, tearing paper
Musical instruments	1079	drum, whistle, bugle
Electronic sound	705	phone, toy
Mechanical	1000	spring, stapler

3.1. Design of environmental sound HMMs and speech HMMs

We first conducted the training of the environmental sound HMMs for the environmental sound source identification. The HMMs were left-to-right models with three states, and the feature vectors employed 16 orders of MFCC, 16 orders of Δ MFCC and one order of Δ power. The sampling frequency was 12 kHz. Table 2 shows the experimental condition. We conducted two evaluation experiments: Exp. A and Exp. B. Exp. A is the identification with environmental sound sources, and Exp. B is the classification with environmental sound sources and speech.

3.2. Experimental results for the identification with environmental sound sources (Exp. A)

The environmental sound HMMs were trained with 92 kinds of environmental sounds \times 45 samples. We first evaluate the environmental sound source identification with the single-occurrence environmental sounds. 92 kinds of single-occurrence environmental sounds \times 5 sets are used as the test data (NSET). We also evaluate the environmental sound source identification with the continuous-occurrence environmental sounds. 92 kinds of continuous-occurrence environmental sounds \times 5 sets are used as the test data (CNSET). The continuous-occurrence environmental sounds are designed by connecting the single-occurrence environmental sounds. Table 3 shows the experimental results in the above experimental condition. Based on the evaluation experiments, this approach can achieve a higher identification performance in single-occurrence environmental sound source identification. Figure 1 shows examples of the spectrum of the identification error (clipping sound and breaking stick sound) in this evaluation experiment.

Compared with the results of the single-occurrence environmental sound source identification, the continuous-occurrence environmental sound source identification performance is significantly degraded. This is because one continuous-occurrence environmental sound will be identified as two or more similar single- or continuous-occurrence environmental sounds. Figure 2 shows examples of the spectrum of the identification error (bottle (glass)

Table 2: Experimental condition

Environmental sound HMMs and Speech HMMs	
Num. of states	3 states
Feature vector	MFCC Δ MFCC Δ Power
Sampling freq.	12 kHz
Env. sound DB	RWCP sound scene DB [5]
Speech DB	ATR Speech DB [6]
Exp. A (Identification with environmental sound sources)	
Num. of models	92 models
Training data	92 kinds of env. sounds \times 45 samples
Test data	Open test
NSET	92 kinds of single-occurrence env. sounds \times 5 sets
CNSET	92 kinds of continuous-occurrence env. sounds \times 5 sets
Exp. B (Classification with environmental sounds and speech)	
Num. of models	Env.: 1 model Speech: 1 model
Training data	Env.: 92 kinds of env. sounds \times 20 samples Speech: 2620 Japanese words (one speaker)
Test data	Open test
SNSET	{92 kinds of single-occurrence env. sounds + 216 Japanese words (speech)} \times 5 sets
SCNSET	{92 kinds of continuous-occurrence env. sounds + 216 Japanese words (speech)} \times 5 sets

Table 3: Experimental results for env. sound source identification

Test data	Single-occurrence	Continuous-occurrence	
	Ident. rate[%]	Test data	Ident. rate[%]
NSET1	96.7	CNSET1	69.5
NSET2	94.6	CNSET2	67.4
NSET3	96.7	CNSET3	72.6
NSET4	94.6	CNSET4	60.0
NSET5	94.6	CNSET5	60.0
Ave.	95.4	Ave.	65.9

clink and cup clink) in this evaluation experiment.

Although this approach did not achieve a higher identification performance in continuous-occurrence environmental sound source identification, we can confirm that the identification errors tend to occur in a similar category as shown in Table 4. Therefore, this approach promises to cluster the environmental sound sources to some categories with clustering models of environmental sound sources for robust speech recognition.

3.3. Experimental result for classification with environmental sound sources and speech (Exp. B)

Next, we tried to conduct a two-class classification between environmental sound sources and speech with the HMM. We designed one environmental sound model and one speech model for the classification. The environmental sound model was trained with 92 kinds of environmental sounds \times 20 samples and the speech model was trained with 2620 Japanese words.

We evaluate the two-class classification performance by using the 216 Japanese isolated words and 92 kinds of single-occurrence environmental sounds \times 5 sets (SNSET), and the 216 Japanese isolated words and 92 kinds of continuous-occurrence environ-

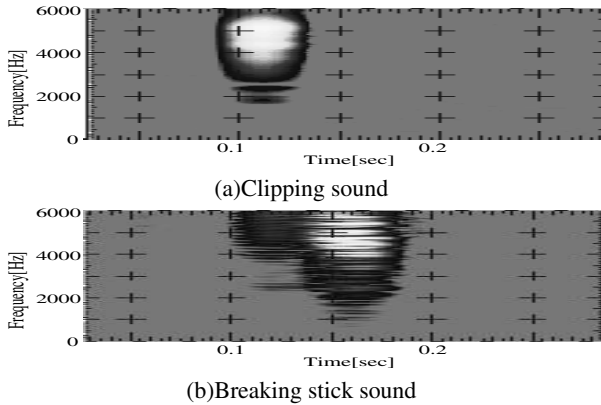


Figure 1: An example of the spectrum of identification error results in single-occurrence env. sound source identification

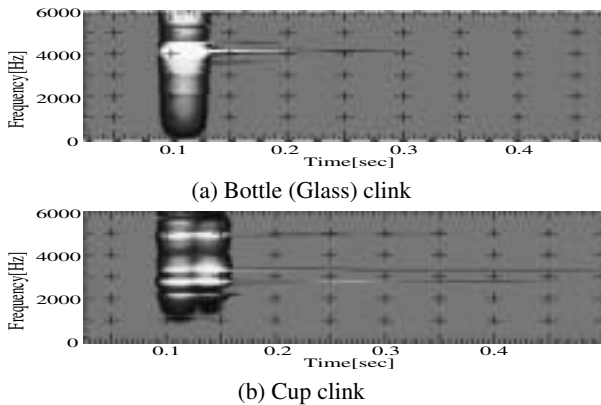


Figure 2: An example of the spectrum of identification error results in continuous-occurrence env. sound source identification

mental sounds \times 5 sets (SCNSET) as test data. Table 5 shows the experimental results in the above condition. As a result, we can confirm that the classification of single- and continuous-occurrence environmental sound sources and speech achieves very effective performance. Therefore, this method will be very useful for voice activity detection for robust speech recognition or talker localization for microphone array steering.

4. ENVIRONMENTAL SOUND-ADDED SPEECH RECOGNITION WITH HMM COMPOSITION

We attempted to improve the ASR performance in environmental sound-added speech with a new HMM composition. A conventional HMM composition method composed of speech HMMs and an environmental sound HMM trained using noise periods prior to the target speech in a captured signal. Although this training method is effective in a stationary noisy environment, it is not so effective in a non-stationary noisy environment, such as a real room, because of noise mismatch. Therefore, we propose a new HMM composition method such that the environmental sounds are clustered into a similar category in advance, and the category is selected based on the identification of environmental sound periods prior to the target speech in the captured signal. Then, the categorized environmental sound HMM for the selected category is composed to speech HMMs.

Table 4: An example of identification error results in continuous-occurrence environmental sound source identification

Correct	Identification result
Bottle clink (continuous)	Bottle clink (single) and cup clink (continuous)
Bottle clink (continuous)	Bottle clink (single) and pottery clink (continuous)
Dropping sound of coins (continuous)	Dropping sound of coins (single) and throwing sound of dice (continuous)

Table 5: Experimental results for the two-class classification between env. sound source and speech

Single-occurrence		Continuous-occurrence	
Test data	Class. rate[%]	Test data	Class. rate[%]
SNSET1	99.7	SCNSET1	99.7
SNSET2	100.0	SCNSET2	99.7
SNSET3	99.7	SCNSET3	99.7
SNSET4	100.0	SCNSET4	99.7
SNSET5	99.7	SCNSET5	100.0
Ave.	99.9	Ave.	99.8

4.1. Experimental condition

Table 6 shows the constructional condition of the environmental sound HMM and speech (phoneme) HMM. In the ASR, the speaker dependent Japanese phoneme-balanced isolated 216 words are employed as the test data. We conducted two experiments. One is the single-occurrence bell sound-added speech recognition. In this experiment, the single-occurrence bell sound-added speech is designed by adding the single-occurrence bell sounds (10 samples of bells3 in the RWCP-DB) to speech. The other is the continuous-occurrence bell sound-added speech recognition. In this experiment, the continuous-occurrence bell sound-added speech is designed by adding the continuous-occurrence bell sounds (30 samples of bells1, bells2, and bells3 in the RWCP-DB) to speech. Figure 3 shows the clean speech and continuous-occurrence bell sound-added speech. These test data do not include data while the HMMs are trained. Also, the SNR (Signal-to-noise ratio) is 5 dB in continuous-occurrence bell sound-added speech. Although stationary noise usually does not occur so much, similar noise sound continuously occurs in real acoustic environments (for example, the sound of a door closing). We assume this situation, and simulate it by continuous-occurrence bell sound-added speech with 30 samples of bells1, bells2 and bells3 in the RWCP-DB.

Table 6: Constructional condition of HMMs

Feature vector	MFCC
Sampling freq.	12kHz
Environmental sound HMM	
Num. of states	2 states ergodic
Database	RWCP sound scene database
Phoneme HMM	
Num. of states	3 states
Database	ATR Japanese speech database
Training data	Speaker dependent 2620 Japanese words

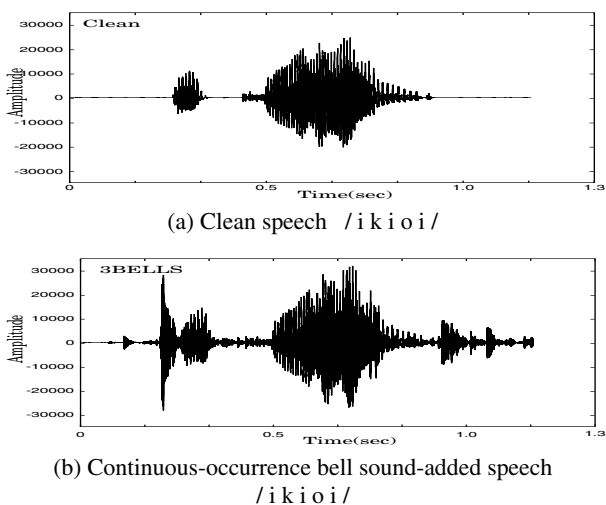


Figure 3: An example of test data on the ASR with HMM composition

4.2. Experimental results

We carried out evaluation experiments on the ASR with the clean speech HMM, conventional HMM composition by one kind of environmental-sound trained HMM and speech HMM, and the proposed HMM composition by the categorized environmental sound HMM for a selected category and speech HMM.

Tables 7 and 8 show the experimental results. Although the conventional HMM composition and proposed HMM composition performances are almost the same in the single-occurrence bell sound-added speech, the proposed HMM composition with the categorized environmental sound HMM for a selected category and speech HMM is about 10% more effective than the conventional HMM composition with one kind of environmental-sound trained HMM and speech HMM.

In the above evaluation experiments, we confirmed that the proposed HMM composition by the categorized environmental sound HMM for a selected category and clean speech HMM achieves effective and robust performance in environmental sound-added speech.

5. CONCLUSIONS

In this paper, we attempted to accurately identify and recognize environmental sound sources for robust speech recognition. As a result, continuous-occurrence environmental sound identification could not achieve effective performance, although single-occurrence environmental sound identification sufficiently achieved the effective performance. However, we confirmed that identification error tends to occur in similar categories in the evaluation experiments. Also, as a result of the classification experiment of environmental sound source and speech, we confirmed that environmental sound source and speech can be classified effectively with an HMM. In addition, we proposed a new HMM composition method with a categorized environmental sound HMM and speech HMM. As a result of the evaluation experiment, we confirmed that the proposed HMM composition achieves effective and robust performance in environmental sound-added speech. In future work, we will attempt to automatically classify the environmental sounds

Table 7: ASR results in single-occurrence bells1 sound-added speech

HMM	Recog. rate [%]
Speech HMM	26.4
Conventional HMM composition with bells1 sounds and speech	95.8
Proposed HMM composition with categorized bell sounds and speech	94.4

Table 8: ASR results in continuous-occurrence bells1, bells2, and bells3 sound-added speech

HMM	Recog. rate [%]
Speech HMM	41.2
Conventional HMM composition with bells1 sounds and speech	43.1
with bells2 sounds and speech	77.3
with bells3 sounds and speech	85.2
Proposed HMM composition with categorized bell sounds and speech	95.8

into categories based on the above proposed method for robust speech recognition.

6. ACKNOWLEDGEMENT

This research was partially supported by The Telecommunications Advancement Organization of Japan and The Ministry of Education, Culture, Sports, Science and Technology of Japan under Grant-in-Aid No. 14780288.

7. REFERENCES

- [1] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-steered Microphone Arrays for Sound Transduction in Large Rooms," J. Acoust. Soc. Am., Vol. 78, No. 5, pp. 1508–1518, Nov. 1985.
- [2] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. ASSP, Vol. ASSP-27, No. 2, pp. 113–120, 1979.
- [3] F. Martin, K. Shikano, and Y. Minami, "Recognition of Noisy Speech by Composition of Hidden Markov Models," Proc. EUROSPEECH'93, pp. 1031–1034, Sep. 1993.
- [4] T. Takiguchi, S. Nakamura, and K. Shikano, "HMM-Separation-Based Speech Recognition for a Distant Moving Speaker", IEEE Trans. SAP, Vol. 9, No. 2, pp.127–140, Feb. 2001.
- [5] S. Nakamura, K. Hiyane, F. Asano, T. Yamada, and T. Endo, "Data Collection in Real Acoustical Environments for Sound Scene Understanding and Hands-Free Speech Recognition," Proc. EUROSPEECH'99, pp. 2255–2258, Sep. 1999.
- [6] K. Takeda, Y. Sagisaka, and S. Katagiri, "Acoustic-Phonetic Labels in a Japanese Speech Database," Proc. European Conference on Speech Technology, Vol. 2, pp. 13–16, Oct. 1987.