DESIGN AND COLLECTION OF ACOUSTIC SOUND DATA FOR HANDS-FREE SPEECH RECOGNITION AND SOUND SCENE UNDERSTANDING

Satoshi Nakamura[1], *Kazuo Hiyane*[2], *Futoshi Asano*[3], *Yutaka Kaneda*[4],
Takeshi Yamada[5], *Takanobu Nishiura*[1,6], *Tetsunori Kobayashi*[7], *Shiro Ise*[8], *Hiroshi Saruwatari*[9],

[1] ATR Spoken Language Translation Laboratories 2-2, Hikaridai Seikacho Kyoto 619-0288, Japan
[2]Mitsubishi Research Institute, [3] National Institute of Advanced Industrial Science and Technology,
[4]Tokyo Denki University, [5]Tsukuba University, [6]Wakayama University,
[7]Waseda University, [8]Kyoto University, [9]Nara Institute of Science and Technology

ABSTRACT

The sound data for open evaluation is necessary for the studies such as sound source localization, sound retrieval, sound recognition and hands-free speech recognition in real acoustic environments. This paper reports on our project aiming the acoustic data collection. There are many kinds of sound scenes in real environments. The sound scene is specified by sound sources and room acoustics. The number of combination of the sound sources, source positions and rooms is huge in real acoustic environments. We assumed that the sound in the environments can be simulated by convolution of the isolated sound sources and impulse responses. As an isolated sound source, hundred kinds of environment sounds and speech sounds are collected. The impulse responses are collected in various acoustic environments. Additionally we collected sounds from the moving source. In this paper, progress of our sound scene database collection project and application to environment sound recognition and hands-free speech recognition are described.

1. INTRODUCTION

Human beings really sense the surrounding environments accurately integrating both visual and auditory information complementary. These kinds of information are essential for human interaction with the environment. For instance, the auditory information plays a more important role for sensing the rear environments. Here, we call the sound environments by the word *sound scene*.

Almost all research on auditory information has been conducted focusing on the individual study of acoustic signal processing, auditory processing, and speech communication. However, the most important point is the close cooperation and integration of these functions to understand the sound scene. To understand a specific sound, the system needs to localize the target sound among multiple sound mixtures in the environment, and focus on the sound.

Recently importance of hands-free speech communication is increasingly recognized. The hands-free speech recognition will bring us so natural and friendly man-machine interface that users are not encumbered by microphone equipments and that users can utter from distance while moving. This hands-free speech recognition is actually an urgent technology for the hands-free interface of a car navigation system and a cellular telephone in the car. If the speaker utters the speech from distance, the accuracy will be seriously degraded by the influences of the noise and reverberation of the room (fig.1). The speech recognition performance even using a desktop microphone will be also varied if the distance between a mouth and a microphone is changed, and if the speaker turns his face to another direction. The fundamental problems of hands-free recognition already have lain in the previous speech recognition framework. To these problems, the following technologies are required,

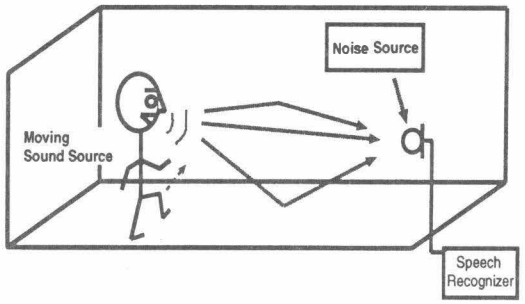

Figure 1: Real environments

- Robustness to directional noise and omni-directional noise (diffuse noise) in the room.
- Robustness to acoustic reflection and reverberation in the room.
- Localization, tracing and recognition of the speaker among many sound sources including other speakers and noise.

These problems are quite new ones which previous studies haven't been considered. In fact, performance of current LVCSR(Large Vocabulary Continuous Speech Recognition) will be seriously degraded if used in this hands-free context.

To conduct these researches, the collection of sound scene data in real acoustic environments is indispensable. The sound scene database contributes to promote a study of sound scene understanding. Only a few databases were developed for the study of sound mixtures. ShATR [1], reported in 1994, is a database of multi-simultaneous-speakers. Spoken dialogues of five speakers using five headset microphones and one desktop microphone were collected. Video images are also recorded by a camera mounted at the ceiling. However, the ShATR focused only on a study of human perception of mixture of speech utterances in natural surroundings. On the other hand, CAIP and IRST reported databases collected using a microphone array in [2, 3, 4]. These databasesare

0-7803-7304-9/02/$17.00 ©2002 IEEE

Table 1: Dry Source sound

	Category	#samples	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

very valuable for the microphone array studies. However, the variety of acoustic environments is very limited for a study of sound scenes in real acoustic environments.

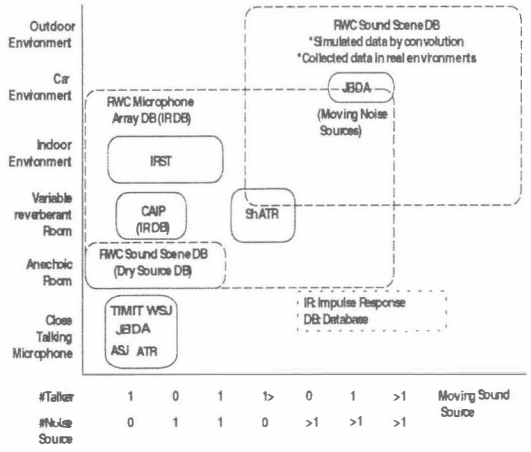

Figure 2: Focus of the RWCP sound scene database from a point of view of sound sources and acoustic environments

Figure 2 shows the focus of the RWCP(Real World Computing Partnership) sound scene database from the point of view of sound sources and acoustic environments. JEIDA database [5], ATR database [6], and ASJ database [7] are databases collected only for study of speech recognition using a close talking microphone. JEIDA also includes noise data collected in a car while driving on the real road. In the figure, the project aims to cover the real acoustical environments including multiple talkers and speakers, indicated as RWC Sound Scene DB. This is a result of combination of RWC Microphone Array DB of many impulse responses measured in various acoustical environments and RWC Sound Scene DB of various dry sound sources collected in the anechoic room. As indicated in the figure, the RWCP sound scene database aims to collect a variety of sound scenes systematically. The figure also indicates the database for the study of source localization, sound retrieval, sound recognition and speech recognition for hands-free speech communication and security systems is necessary. In this paper, we describe our sound scene database which is composed of isolated environment sounds and impulse responses in various rooms. Then the results of the isolated environment sound recognition experiments are also described[8, 9].

2. RWCP SOUND SCENE DATABASE

This project is one of the projects supported by RWCP (Real World Computing Partnership). The objective of the project is to provide common standard databases for research concerning real acoustic environments.

It is almost impossible to collect all combinations of the existing sound sources and real acoustic environments. Thus, we started to collect two kinds of sound data. The first data is isolated sounds of environment non-speech sounds and speech sounds. We call the isolated sounds recorded in an anechoic room by the word *dry source* in this paper. The dry source is free from influences of room acoustics. The second data is impulse responses in various acoustic environments. The sound in the environment can be simulated by convolution of the dry sources and the impulse responses. However, there are sounds which is unable to simulate by the convolution such as non point source sounds and moving sound sources. We collected those sounds using a three dimensional microphone array. The microphone array database enables to extract arbitrary sounds by various beam-forming algorithms.

The data is collected in an anechoic room, a variable reverberant room, office environments, where many sound sources exist. Various kinds of sound sources including speech are also collected as target sounds.

3. DATA COLLECTION

3.1. Dry Source Database

Dry source is the sound recorded in an anechoic room which is free from room acoustics. The environment sound can be simulated by convolution of the dry source and an impulse response if the

transmission channel is linear and stationary. We collected three kinds of environment sounds shown in table 1. The first class is collision sounds of wood, plastic and ceramics. The second class and the third class are composed of sounds occurred when human beings operate on things like spray, saw, claps, coins, books, pipes, telephones, toys, etc. The sounds of the second class are the sounds whose source materials can not be easily associated. Whereas the source materials of the third class sounds can be easily associated uniquely.

We recorded around 100 samples for about 90 kinds of sounds sufficient enough for statistical model training. The recording is conducted in an anechoic room by B&K 4134 microphone and DAT recorder in 48kHz 16bit sampling. SNRs of the data are around 40-50dB.

3.2. Impulse Response Database

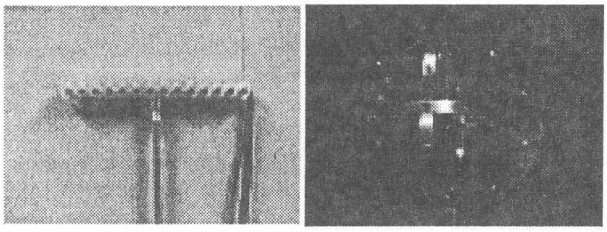

Figure 3: 14ch linear and 54ch spherical microphone arrays

Table 2: Recording conditions for impulse responses

A/D, D/A	Pavec MD-8000mk2 64ch 24bit
Microphone	54ch Spherical array 14ch Linear array 16ch Circle array
Source	Diatone DS-7 loud speaker B&K Type 4128 Head-Torso
Source Sounds	Time stretched pulse Balanced words(216) Balanced sentences: TIMIT SX(40), ATR(50)

We collected impulse responses at different locations in different rooms. The sounds are recorded in an anechoic room, a variable reverberant room and offices using 3 kinds of microphone arrays by the Diatone DS-7 loud speaker and B&K Type4128 Head-Torso. Reverberation times of the rooms are 9 variations from 0.0 to 1.3 seconds. Table 2 shows recording conditions of impulse responses. Fig.3 shows a 14ch linear microphone array and a 54ch spherical microphone array used in the data collection.

Fig.4 shows a variable reverberant room whose reverberation time can be adjusted from 0.3 to 1.3 seconds by changing reflection walls. The impulse responses are measured from different angles from the sound source and a microphone.

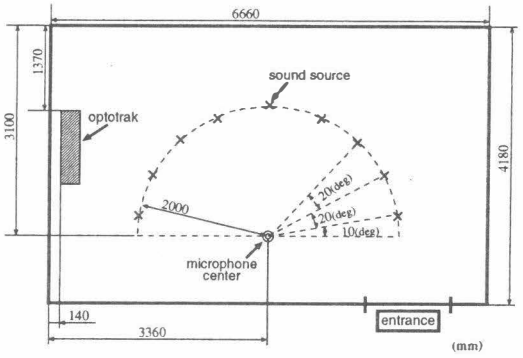

Figure 4: Data collection in a variable reverberation room

3.3. Moving Sound Source

The sound in the real room can be simulated by convolution only if the transmission channel is linear and stationary. However, speakers may move while uttering in the real situation. We collected a moving sound source with respective position (x,y,z) simultaneously by OPTOTRAK. The OPTOTRAK is an infrared optical position sensing system with very high position resolution whose RMS resolution is 0.1mm. Phonetically balanced words and sentences (TIMIT SX sentence set and ATR balanced sentence set) are played through a loud speaker attached to moving sound system. Fig.5 and fig.6 show the moving sound system we developed and an example moving sound source position trajectory for a sentence utterance.

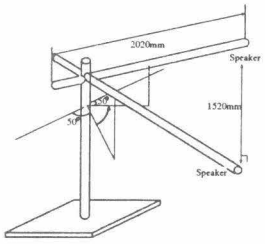

Figure 5: Equipment used for moving sound data collection

4. ENVIRONMENT SOUND RECOGNITION BASED ON HMMS

The collected dry source data is available to use as a typical environment sound in the real environments. The objectives of our research are sound scene understanding and hands-free speech recognition in the environments where many sounds exist. We try to evaluate environment sound recognition by the hidden Markov models. Specifications of HMMs are depicted in the table 3. The environment sound recognition experiments are carried out for the single occurrence and the multiple occurrence of the same environment sounds. Five test sets both for single and multiple occurrences are evaluated. Table 4 are results for the experiments.

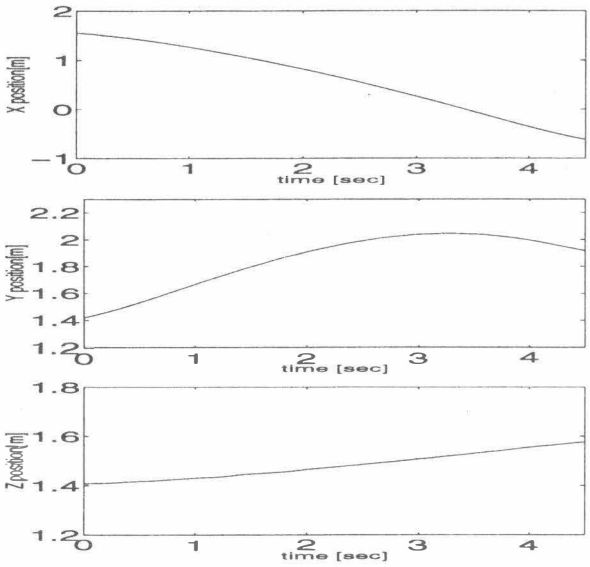

Figure 6: Position trajectory for the moving sound

Table 3: Condition of HMMs

Sampling	12kHz, 16bit
Features	MFCC, ΔMFCC,ΔPower
#HMM states	1,2,3 states
PDF	Mixture Gaussian density 1-20 pdfs
Training Set	42 samples for 92 kinds of environment sounds
Test Set	20 samples for 92 kinds of environment sounds

Since the average rate of 95.4% is very high, the recognition system with the feature extraction and HMMs successfully recognize the collected environment sounds. For the multiple occurrence of the environment sounds, the average rate of 88.7% is relatively lower than that of the single occurrence though, the rate is still very high. This results confirm that the statistical modelling is very effective not only for speech recognition but also for the environment sounds recognition, if sufficient number of training data can be prepared. We are currently applying the modelling of environment sounds to sound source identification.

Table 4: Recognition Accuracy[%]

Single Set-1	96.7	Multiple Set-1	91.3
Single Set-2	94.6	Multiple Set-2	88.0
Single Set-3	96.7	Multiple Set-3	92.4
Single Set-4	94.6	Multiple Set-4	85.9
Single Set-5	94.6	Multiple Set-5	85.9
Single Ave.	95.4	Multiple Ave.	88.7

5. CONCLUSION

This paper describes sound scene data collection indispensable for studies of sound understanding including sound source localization, sound retrieval, sound recognition and speech recognition in real acoustic environments. We collected a dry source database and an impulse response database. Furthermore, we collected a moving sound data with the respective sound source position, since moving sound source can not be simulated by the convolution. Then we tried to recognize the collected environment sound using a dry source database based on HMMs.

The collected data is scheduled to be distributed freely for research purposes by three DVD-ROMs containing the acoustic dry sound source data, impulse responses, and sound position information. The information regarding to this database is summarized in the following URL:

http://tosa.mri.co.jp/sounddb/indexe.htm

The page also introduces our schedule of distribution and a way to get the DVDs. The URL includes not only the database specification but also theories and methods of measurement of an impulse response by TSP, estimation of reverberation time, sound source localization, and convolution. The application researches using the RWCP database are also described.

6. ACKNOWLEDGEMENT

This research was supported in part by the Telecommunications Advancedment Organization of Japan

7. REFERENCES

[1] M. Crawford, G. J. Brown, M. Cook, and P. Green, "Design, collection and analysis of multi-simultaneous-speaker corpus," *Proc. the Institute of Acoustics, Vol.16, Part 5*, pp. pp.183–190, 1994.

[2] Q. Lin, C. Che, and J. French, "Description of the caip speech corpus," *Proc. ICSLP*, 1994.

[3] E. Jan, P. Svaizer, and J. Flanagan, "A database for microphone array experimentation," *Proc. Eurospeech*, 1995.

[4] D. Gluliani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of different microphone array configurations for hands-free speech recognition in noisy and reverberant environment," *Proc. Eurospeech*, 1997.

[5] S. Itahashi, "Recent speech database projects in japan," *Proc. ICSLP*, 1990.

[6] K. Takeda, Y. Sagisaka, S. Katagiri, and H. Kuwabara, "A japanese speech database for various kinds of research purposes," *Proc. ICSLP*, 1988.

[7] T. Kobayashi, S. Itahashi, and T. Takezawa, "Asj continuous speech corpus for research," *Journal of Acoustical Society of Japan*, pp. 48. 12. pp.888–893, 1992.

[8] 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. Eurospeech99*, 1999.

[9] S. Nakamura, K. Hiyane, F. Asano, T.Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *Proc. ELREC2000*, 2000.