Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition

Satoshi Nakamura[1], Kazuo Hiyane[2], Futoshi Asano[3],
Takanobu Nishiura[1,4], Takeshi Yamada[5]

[1] ATR Spoken Language Translation Labs. 2-2, Hikaridai Seikacho Kyoto 619-0288, Japan, nakamura@slt.atr.co.jp
[2]Mitsubishi Research Institute 2-3-6, Otemachi Chiyoda Tokyo 100-8141, Japan, hiya@mri.co.jp
[3] Electrotechnical Laboratory 1-1-4, Umezono Tsukuba Ibaraki 305, Japan, asano@etl.go.jp
[4] Nara Institute of Science and Technology 8916-5 Takayama Ikoma Nara 630-0101, Japan, takano-n@is.aist-nara.ac.jp
[5]Tsukuba University 1-1-1, Tennodai Tsukuba Ibaraki 305, Japan, takeshi@is.tsukuba.ac.jp

Abstract

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

1. Introduction

Generally, auditory as well as visual information is quite important for human beings to sense surrounding environments. This information is essential for human interaction with the environment. Human beings really sense the surrounding environments accurately integrating both visual and auditory information complementary. 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 acoustical 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. To conduct the research of the sound scene, the collection of sound scene data in real acoustical 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(Crawford et al., 1994), 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 (Lin et al., 1994; Jan et al., 1995; GIuliani et al., 1997). These databases are very valuable for the microphone array studies. However, the variety of acoustical environments is very limited for a study of sound scenes in real acoustical environments.

Figure 1 shows the focus of the RWCP sound scene database from the point of view of sound sources and acoustical environments. JEIDA database(Itahashi, 1990), ATR database(Takeda et al., 1988), and ASJ database(Kobayashi et al., 1992) 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. As indicated in the figure, the RWCP sound scene database aims to collect a variety of sound scenes systematically. The figure also indicates the lack of the database for the study of source localization, sound retrieval, sound recognition and speech recognition for hands-free speech communication and security systems.

Figure 1: Focus of the RWCP sound scene database from a point of view of sound sources and acoustical environments

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.

2. 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 acoustical environments.

It is almost impossible to collect all combinations of the existing sound sources and real acoustical 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 acoustical 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 are planning to collect those sounds using a three dimensional microphone array. The microphone array database enables to extract arbitrary sounds by various beamforming 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 stable. We collected three kinds of environment sounds shown in Table **??**. The first class is crash 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

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 from 0.01 to 2 seconds. Table 2 shows recording conditions of impulse responses.

Table 1: Source sound

Class 1	Wood	Wooden boads and a wooden stick
	Metal	Metal boads and a metal stick
	Plastic	Plastic boads and a plastic stick
Class 2	Friction Noise	Sound of the saw
	Prosive Noise	Break chop sticks
	Burst Noise	Claps
Class 3	Metals	Bells
	Music instruments	Whistles
	Electronic sounds	Telephone Rings

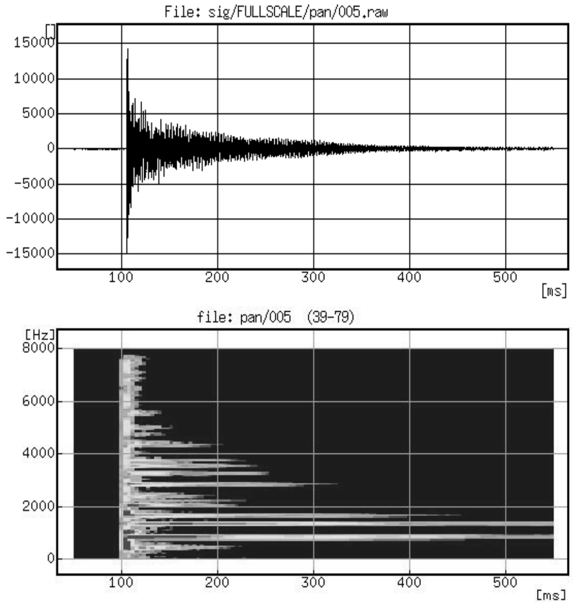

Figure 2: Waveform and spectrogram of dry sound hitting a metal board by a wooden stick

Table 2: Recording conditions for impulse responses

A/D, D/A	Pavec MD-8000mk2 64ch 24bit
Microphone	54ch Spherical array 14ch Linear array(2.83cm spacing) 16ch Circle array
Source	Diatone DS-7 loud speaker B&K Type 4128 Head-Torso
Source Sounds	Time stretched pulse Phonetically balanced words(216) Phonetically balanced sentences (TIMIT(40), JNUS(50)) Real speech

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. Fig.5 shows an impulse response and its frequency response measured in the variable reverberation room where its reverberation time is 1.3 second.

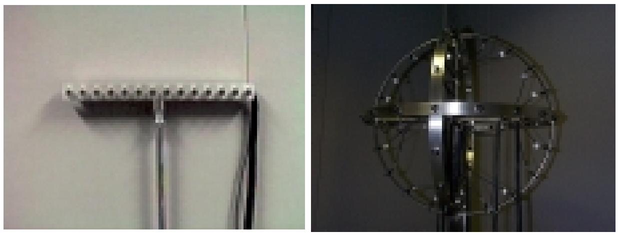

Figure 3: 14ch linear and 54ch spherical microphone arrays

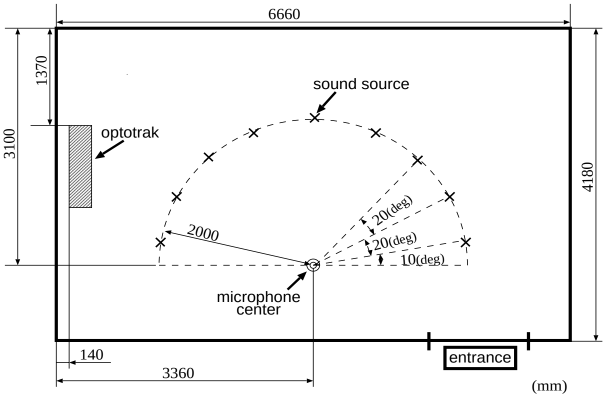

Figure 4: Data collection in a variable reverberation room

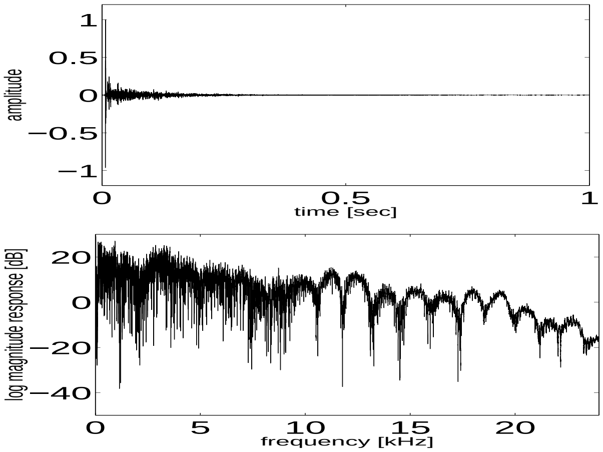

Figure 5: Impulse response and frequency response in a variable reverberation room (T60=1.3sec.)

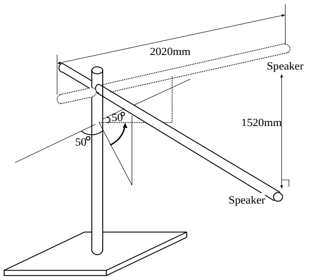

Figure 6: Equipment used for moving sound data collection

3.3. Moving Sound Source

The sound in the real room can be simulated by convolution only if the transmission channel is linear and stable. 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 sensering system with very high position resolution whose RMS resolution is 0.1mm. Phonetically balanced words and sentences are played through a loud speaker attached to moving sound system. Fig.6 and Fig.7 show the moving sound system we developed and an example moving sound source position trajectory for a sentence utterance.

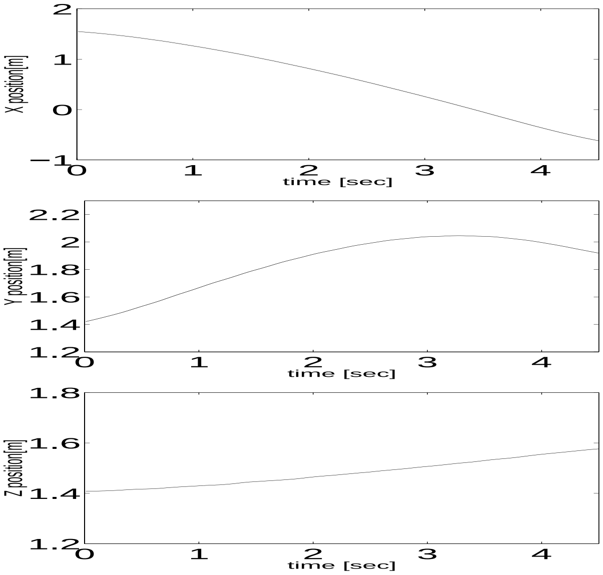

Figure 7: Position trajectory for the moving sound

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 exit. We try to evaluate environment sound recognition by the hidden Markov models. Specifications of HMMs are depicted in Table 3. The environment 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

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

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. 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 modeling is very effective not only for speech recognition but also for the environment sounds recognition, if sufficient number of training data can be prepared.

5. Conclusion

This paper describes a sound scene data collection project indispensable for studies of sound understanding including sound source localization, sound retrieval, sound recognition and speech recognition in real acoustical environments. Since sounds in the collected environments can be simulated by convolution of a source sound and an impulse response as far as the point source assumption is satisfied. Thus 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. The other sounds in real acoustical environments will be collected directly using a three dimensional data collection system. Then we tried to recognize the collected environment sound using a dry source database based on HMMs. It is confirmed that HMMs is also effective not only for speech recognition but also for environment sound recognition, if we have a sufficient number of training data. The collected data will be distributed freely on CD-ROMs containing the acoustic sound data and images, sound position information. Tagging information and their handling tools will be provided in a future work.

6. References

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

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

Itahashi, S., 1990. Recent speech database projects in japan. *Proc. ICSLP*.

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

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

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

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