J. Acoust. Soc. Jpn. (*E*) **20**, 3 (1999)

Sound scene data collection in real acoustical environments

Satoshi Nakamura,*[1] Kazuo Hiyane,*[2] Futoshi Asano,*[3] and Takashi Endo*[4]

*[1] *Nara Institute of Science and Technology, 8916-5, Takayama, Ikoma, 630-0101 Japan*
*[2] *Mitsubishi Research Institute, 2-3-6 Otemachi, Chiyoda-ku, Tokyo, 100-8141 Japan*
*[3] *Electrotechnical Laboratory, 1-1-4, Umezono, Tsukuba, 305-8568 Japan*
*[4] *Real World Computing Partnership, 1-6-1, Takezono, Tsukuba, 305-0032 Japan*

(*Received 7 August 1998*)

This paper describes a sound scene database necessary for studies such as sound source localization, sound retrieval, sound recognition and speech recognition in real acoustical environments. Many speech databases have been collected for speech recognition so far. The statistical modeling of speech based on the collected speech databases realizes a drastic improvement of speech recognition performance. However, there are only a few databases available for sound scene data including non-speech sound in real environments. A sound scene database is obviously necessary for studies of acoustical signal processing and sound recognition. This paper reports on a project for collection of the sound scene database supported by Real World Computing Partnership (RWCP). 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. Two approaches are taken to build the sound scene database in the early stage of the project. The first approach is to collect isolated sound sources of many kinds of non-speech sounds and speech sounds. The second approach is to collect impulse responses in various acoustical environments. The sound in the collected environments can be simulated by convolution of the isolated sound sources and impulse responses. In a later stage, the sound scene data in real acoustical environments is planned to be collected using a three dimensional microphone array. In this paper, the plan and progress of our sound scene database project are described.

Keywords: Sound scene database, Database, Microphone array, Environment sound

PACS number: 43. 72. Ew, 43. 72. Ne, 43. 72. Dv

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. The visual information has been focused on mainly so far, since the visual information provides richer knowledge of environments than that of auditory information. Then only visual information has been focused on in research of robotics, humanoids and self-driving vehicles. However, the importance of the auditory information has began to be noticed recently. 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*.

On the other hand, almost all research on auditory information has been conducted focusing not on the study of sound scene understanding but on the study of acoustical signal processing, auditory processing, and speech communication. There is indeed a lot of research on acoustical signal processing such as sound source localization, beamforming, echo cancelation, speech synthesis, and speech recognition independently. The most important point is that the close cooperation and integration of these functions are necessary 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,[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 recorded also 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. CAIP and IRST reported databases collected using a microphone array in Refs. 2–4). These databases are very valuable for the microphone array studies. However, the variety of acoustical environments in these databases are very limited to be able to study sound scenes in real acoustical environments.

In this paper, we describe our database which aims to collect real sound scenes using a microphone array. A detailed plan and its current status will be discussed.

2. SOUND SCENE DATABASE PROJECT OVERVIEW

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 start to collect sound data using two approaches in an early stage. The first approach is to collect as many isolated sound sources of non-speech sounds and speech sounds as possible. We call the isolated sound source record-

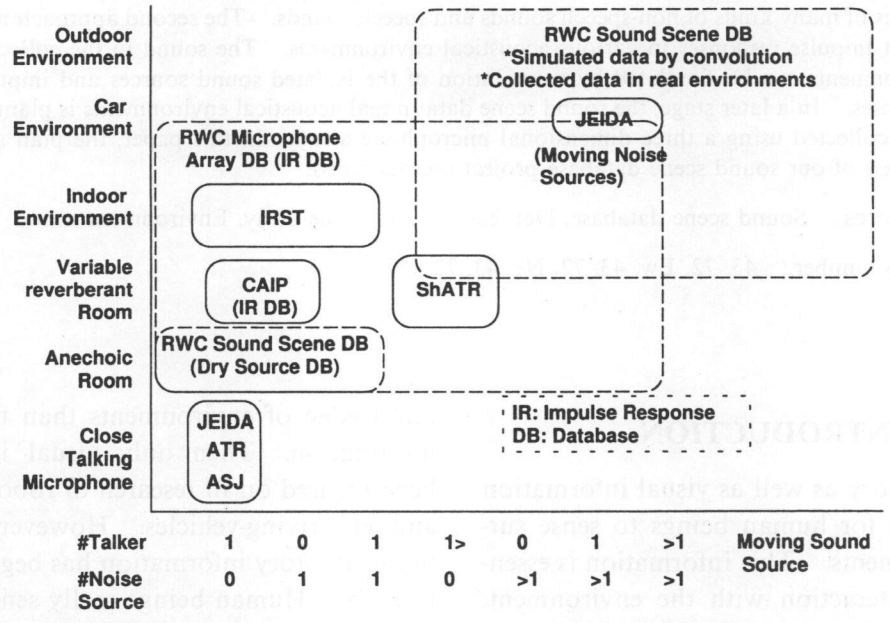


Fig. 1 Focus of the RWCP sound scene database from the point of view of sound sources and acoustical environments.

ed in an anechoic room by the word *dry source* in this paper. The dry source is free from influences of room acoustics. The second approach is to collect impulse responses in various acoustical environments. The sound in the collected environments can be simulated by convolution of the dry sources and the impulse responses.

In a later stage, the sound scene data in real acoustical environments is planned to be collected using a three dimensional microphone array. The microphone array database enables to extract arbitrary sound by various beamforming algorithms.

The database is planned to be collected in an anechoic room, a variable reverberant room, a business office and in outdoor environments where many sound sources exist. Various kinds of sound sources including speech are also planned to be collected as target sounds.

Figure 1 shows the focus of the RWCP sound scene database from the point of view of sound sources and acoustical environments. JEIDA,[5] ATR,[6] and ASJ[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. As indicated in the figure, the RWCP sound scene database aims to collect a variety of sound scenes systematically. Figure 2 shows the focus of the RWCP sound scene database from the point of view of technologies and applications. The figure 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. The figure also clarifies that the database should be collected with three dimensional spatial resolution using a three dimensional microphone array.

2.1 Database

It is necessary to specify the data conditions. Three conditions are to be specified, such as;

- Sound sources (Table 1)
- Environments (Table 2)
- Microphones (Table 3)

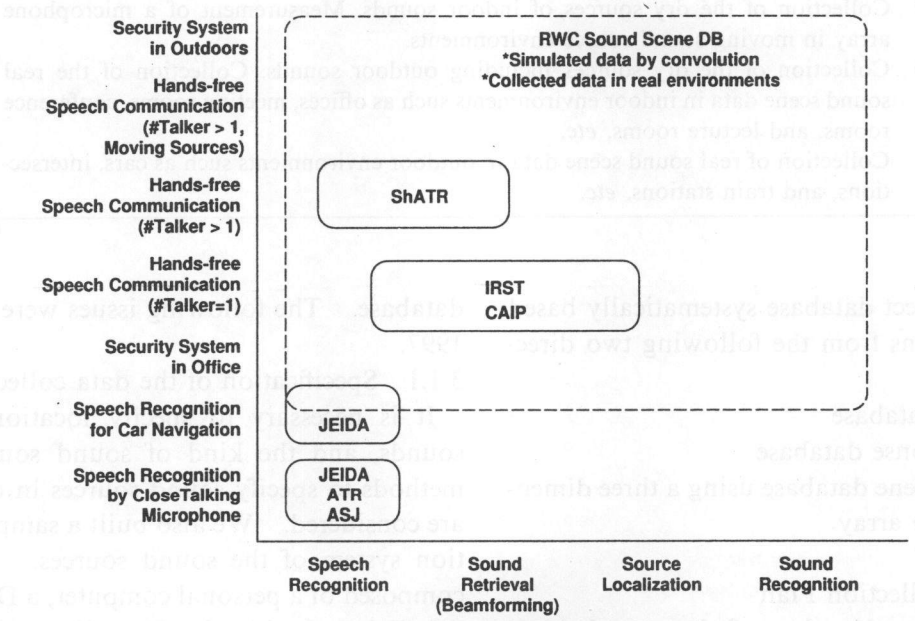


Fig. 2 Focus of the RWCP sound scene database from the point of view of technologies and applications.

Table 1 Source sound.

Speech	Word, Sentence
Background noise	Wind, Rain, Hum of voices, Air conditioner, Omnidirectional noise, Computer noise
Short noise	Crash noise, Friction noise
Long noise	Whistle, Car noise
On off noise	Footstep, Bicycle, Machine noise, Machine gun noise

Table 2 Source location and environments.

Distance	Near, Middle, Far field
Intensity level	Signal to noise ratio (SNR)
Environments	Anechoic room, Reverberant room, Office, Meeting room, Lecture room, Station
Transfer function	Impulse response (TSP, M-random series), Direct-reflection ratio
Sound movement	Direction, Speed, Pattern
Measurement	Real speech, Playback from loud speaker

Table 3 Microphone array.

Microphone	Omni-directional, Uni-directional, Microphone array
Characteristic	Impulse response in anechoic room
Channels	8,16, 64, 112
Spacing	2.83 cm, 3 cm, *etc.*
Array design	Equal Distance, Harmonic, Linear, 2,3-dimensional array

Table 4 Plan for Sound scene database collection.

1997	Planning of the sound scene collection. Collection of sample data. Measurement of a microphone array.
1998	Collection of the dry sources of office sounds in the anechoic room. Measurement of a several designs of a microphone array in the anechoic room and the variable reverberant room.
1999	Collection of the dry sources of indoor sounds. Measurement of a microphone array in moving sound source environments.
2000	Collection of the dry sources including outdoor sounds. Collection of the real sound scene data in indoor environments such as offices, meeting rooms, conference rooms, and lecture rooms, *etc.*
2001	Collection of real sound scene data in outdoor environments such as cars, intersections, and train stations, *etc.*

We plan to collect database systematically based on the specifications from the following two directions.

1. Dry source database
2. Impulse response database
3. Real sound scene database using a three dimensional microphone array

2.2 Database Collection Plan

Table 4 describes the plan of the sound scene database collection. The project is originally scheduled for five years to complete the real sound scene database using a microphone array.

3. CURRENT STATUS

3.1 Dry Source Database

There are many kinds of sound sources in the real sound scene. We started to collect sample sound data of limited kinds of sounds for the dry source database. The following issues were considered in 1997.

3.1.1 Specification of the data collection

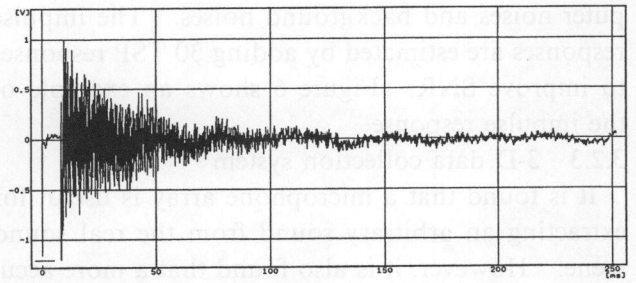
It is necessary to specify location, number of sounds, and the kind of sound sources. Several methods to specify sound sources in the real world are considered. We also built a sample data collection system of the sound sources. The system is composed of a personal computer, a DSP board and a A/D board with a low pass filter and an amplifier. This system is able to record stereo signals. Only monaural signals were collected in 1997.

3.1.2 Collection of single noise

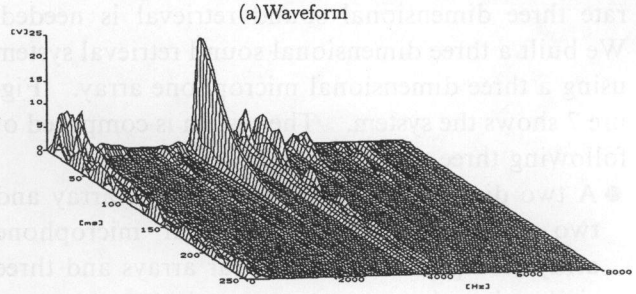
The sound data composed of short sounds, whose duration is about 500 ms, is collected in quiet office environments for consideration of the database design. Fifty kinds of short crash noises, whose duration is about 250 ms, were collected in quiet office environments in 1997. These sounds were

Table 5 Collected sounds.

Sound	Conditions
Crash noise (Wood)	Board+stick
Crash noise (Metal)	Board+stick, Coin drop, Bell ring, Lock
Crash noise (Plastic)	Case strike, Dice drop
Crash noise (Others)	Drawer shut, Clap, Book drop, *etc.*
Burst noise	Firecracker
Frictional noise	Saw sound, Sandpaper
Human noise	Clap, Cough, Cluck, Crunch
Others	Cap open, Paper grasp, *etc.*



(a) Waveform



(b) Spectrogram (0~8kHz, 0~250ms)

Fig. 3 Waveform and spectrogram striking a can by a metal stick.

Table 6 Microphone array system.

Items	specifications
Playback	Loud speaker
M. Array	14 ch, 2.83cm Equi-spaced array
A/D, D/A	16 ch synchronous AD/DA

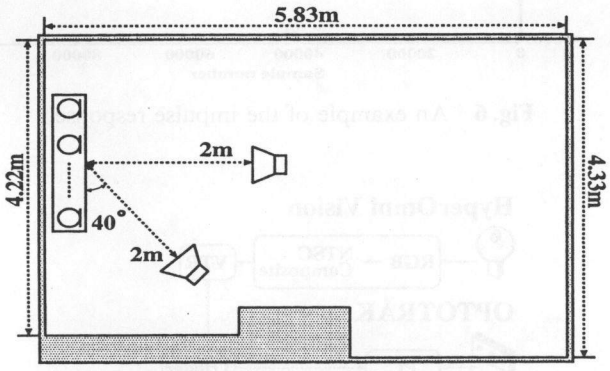


Fig. 4 Sound source location.

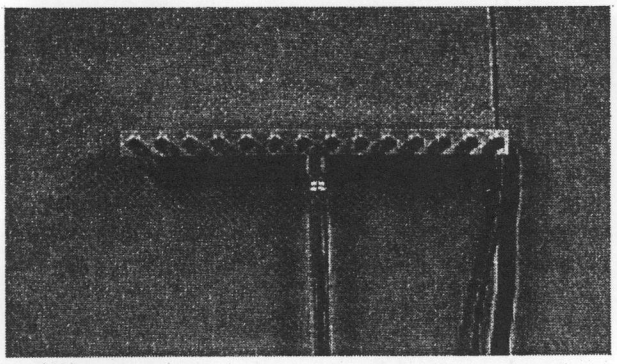


Fig. 5 14 ch microphone array.

selected as representative sounds in the office environments. The sounds to be collected will be extended to the sounds not only in indoor environments but also in outdoor environments. Table 5 shows the collected database. The distance between a sound source and a microphone is about 10–20 cm. The SNR to the background noise is 20–30 dB. Figure 3 shows a waveform and a spectrogram of a sound signal striking a metal can by a metal stick.

3.2 Impulse Response Database and Real Sound Scene Database Using a Microphone Array

We start to measure fundamental characteristics of a microphone array before applying it to the real sound scene. It is necessary to design the microphone array such that it is suitable for the sound scene database. Then various kinds of impulse responses can be collected by the microphone array.

3.2.1 Data collection system

Table 6 shows the set up of the microphone array measurement. The 14 ch microphone array system is used to examine the microphone array characteristics. Figure 4 shows an overview of the experiment

Table 7 Database using a microphone array.

Data	Sound direction	Utter.	Sampling rate	Environment
Phonetically balanced words	90°	216	12 kHz	Experiment room
White Gaussian noise	90/40°	2	〃	〃
Computer noise	90/40°	2	〃	〃
No source	90/40°	1	〃	〃
Impulse response	90/40°	1	48 kHz	Experiment room
Impulse response	90°	1	〃	Sound proof room

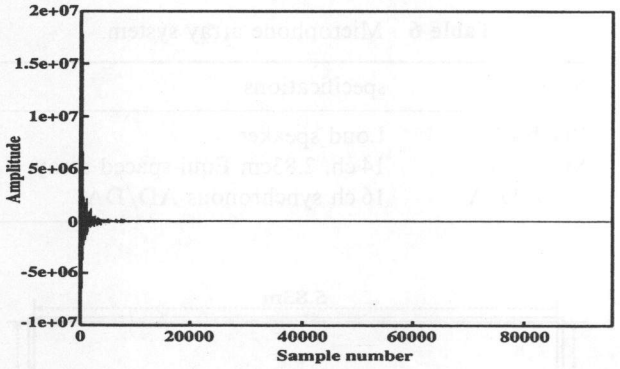


Fig. 6 An example of the impulse response.

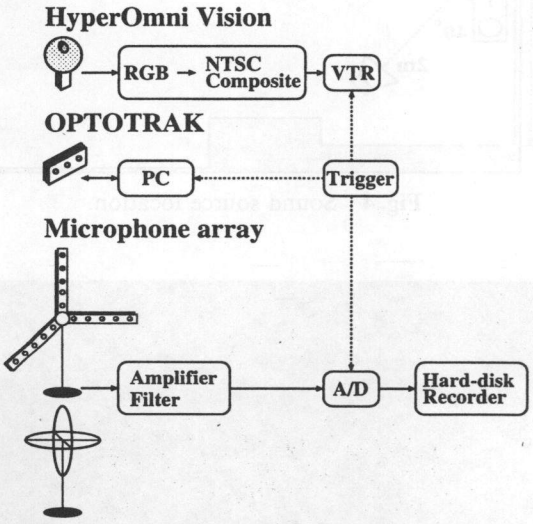


Fig. 7 3-D sound scene data collection system.

room whose reverberation time is about 180 ms. Figure 5 shows the 14 ch microphone array system used in the data collection. TSP signal (time stretched pulse) is used to measure the impulse responses.

3.2.2 Collected sounds

Table 7 shows the data collected for investigation of fundamental characteristics of a microphone array. The data consists of phonetically balanced Japanese 216 words, white Gaussian noise, computer noises and background noises. The impulse responses are estimated by adding 30 TSP responses to improve SNR. Figure 6 shows an example of the impulse response.

3.2.3 3-D data collection system

It is found that a microphone array is useful for extracting an arbitrary sound from the real sound scene. However, it is also found that a more accurate three dimensional sound retrieval is needed. We built a three dimensional sound retrieval system using a three dimensional microphone array. Figure 7 shows the system. The system is composed of following three components.

- A two dimensional circle microphone array and two kinds of three dimensional microphone arrays composed of three linear arrays and three circle microphone arrays, respectively.
- OPTOTRAK, a position sensing system for moving sound sources.
- Hyper omni-vision, a image recording system of 360 angles.

The position sensing system and hyper omni-vision system are necessary for tagging the sound data. We are also planning to prepare handling software tools and automatic tagging tools.

4. 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.

Two approaches were taken in the early stages to build a sound scene database such as a dry source database and an impulse response database. The sound scene data in real acoustical environments is planned to be collected using a three dimensional microphone array in a later stage.

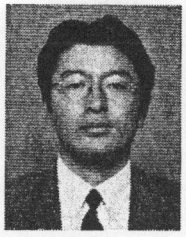
The dry source database is planned to be extended

to include more kinds of sound sources in the anechoic room. This database will be the biggest dry source database of environmental sounds. The impulse response database using a microphone array is planned to be collected in office rooms, conference rooms, lecture rooms, cars, and outdoors with three dimensional spatial resolution. Thus 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. The other sounds in real acoustical environments will be collected directly using a three dimensional data collection system.

The collected data will be distributed freely on CD-ROMs containing the acoustic sound data, tagging information, environment images, sound position informations and their handling tools.

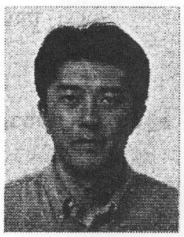
REFERENCES

1) M. Crawford, G. J. Brown, M. Cook, and P. Green, "Design, collection and analysis of a multi-simultaneous-speaker corpus," Proc. Inst. Acoust. **16** (5), 183-190 (1994).
2) Q. Lin, C. Che, and J. French, "Description of the CAIP Speech Corpus," Proc. ICASSP 94, 1823-1826 (1994).
3) E. Jan, P. Svaizer, and J. Flanagan, "A database for microphone array experimentation," Proc. Eurospeech 95, 813-816 (1995).
4) D. Giuliani, 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 97, 347-350 (1997).
5) S. Itahashi, "Recent speech database projects in Japan," Proc. ICSLP 90, 1081-1084 (1990).
6) K. Takeda, Y. Sagisaka, S. Katagiri, and H. Kuwabara, "A Japanese speech database for various kinds of research purposes," J. Acoust. Soc. Jpn. (J) **44**, 747-754 (1988) (in Japanese).
7) T. Kobayashi, S. Itahashi, S. Hayamizu, and T. Takezawa, "ASJ continuous speech corpus for research," J. Acoust. Soc. Jpn. (J) **48**, 888-893 (1992) (in Japanese).



Satoshi Nakamura was born in Japan on August 4, 1958. He received his B.S. degree in electronics engineering from Kyoto Institute of Technology in 1981 and the Ph.D. degree in information science from Kyoto University in 1992. In 1981-1986 and 1990-1993, he worked with Central Research Laboratory, Sharp Corporation, Nara, Japan. From 1986 to 1989, he was a researcher at ATR Interpreting Research Laboratories. In 1996 he was a visiting research professor at CAIP Center of Rugtgers University, USA. He is currently an associate professor at Graduate School of Information Science, Nara Institute of Science and Technology. His current research interests include speech recognition, stochastic modeling of speech and a microphone array. He received the Awaya award from Acoustical Society of Japan, 1992. He is a member of ASJ, IEICE, IPSJ, and IEEE.

Kazuo Hiyane was born in 1963. He received the B.E. and M.E. degrees from Tokyo University in 1986 and 1988, respectively. Since 1988, he has been with Mitsubishi Research Institute (MRI). His research interests include acoustic scene analysis, signal processing, genetic algorithms, and parallel and distributed processing. He is a member of ASJ ; IEICE; and SICE.



Futoshi Asano was born in Fukushima Pref. on October 13, 1962. He received his B.S. degree in electrical engineering, his M.S. and the Ph.D. degrees in electrical and communication engineering from Tohoku University in 1986, 1988, and 1991, respectively. From 1991 to 1995, he was a research associate at R.I.E.C, Tohoku University. From 1993 to 1994, he was a visiting researcher at Pennsylvania State University. Since 1995, he has held the position of researcher at the Electrotechnical Laboratory in Tsukuba. His present research interests include array signal processing, adaptive signal processing, sound field control, and statistical signal processing.

Takashi Endo received his master degree in Engineering from the Waseda University. He was sent on loan to the Real World Computing from 1996 to 1999. He is a researcher of Multimedia System Research Department at the Hitachi Central Research Laboratory. His research interests include speech recognition, speech synthesis, speech signal processing, multi-modal self-organized database and security of microprocessor.