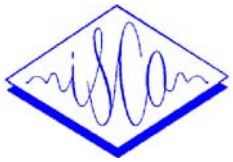


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DATA COLLECTION IN REAL ACOUSTICAL ENVIRONMENTS FOR SOUND SCENE UNDERSTANDING AND HANDS-FREE SPEECH RECOGNITION

[1]Satoshi Nakamura, [2]Kazuo Hiyane, [3]Futoshi Asano, [1]Takeshi Yamada, [4]Takashi Endo
[1]*Nara Institute of Science and Technology, 8916-5, Takayama, Ikoma, Nara, 630-01 Japan,*
[2] *Mitsubishi Research Institute,2-3-6 Otemachi, Chiyoda, Tokyo, 100-8141 Japan,*
[3] *Electrotechnical Laboratory, 1-1-4, Umezono, Tsukuba, Ibaraki, 305 Japan,*
[4] *Real World Computing Partnership, 1-6-1, Takezono, Tsukuba, Ibaraki, 305 Japan,*
Email: nakamura@is.aist-nara.ac.jp, http://www.aist-nara.ac.jp/ñakamura

ABSTRACT

This paper describes a sound scene database necessary for studies such as sound source localization, sound retrieval, sound recognition and hands-free speech recognition in real acoustical environments. This paper reports on a project for collection of the sound scene data 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 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.

1. INTRODUCTION

An 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*.

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 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. CAIP and IRST reported databases collected using a microphone array in [2, 3, 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 are 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 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 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 2. 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.

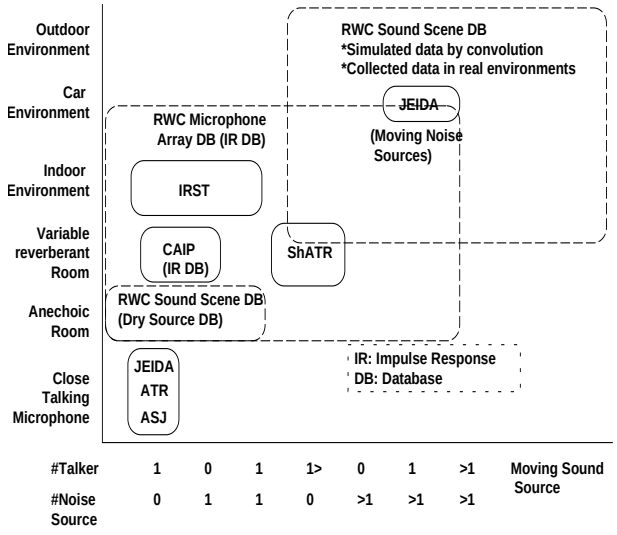


Fig.1 Focus of the RWCP sound scene database from the point of view of sound sources and acoustical environments

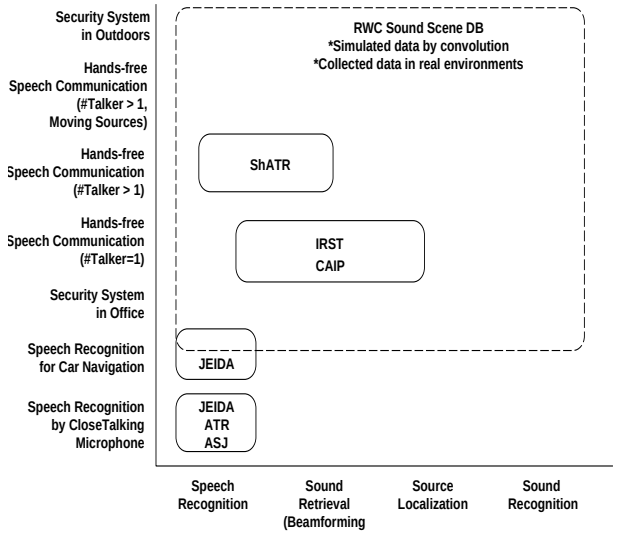


Fig.2 Focus of the RWCP sound scene database from the point of view of technologies and applications

Table 1. Source sound

Speech	Word, Sentence
Background Noise	Wind, Rain, Hum of voices, Air conditioner, Omni directional 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	-10db $\cdots$ +30dB(SNR)
Environments	Anechoic Room, Reverberant Room, Office, Meeting Room, Lecture Room, Station
Transfer Function	Impulse Response (TSP, M-random series),
Sound Movement	Direction, Speed, Pattern
Source Sound	Real Speech, Playback from loud speaker

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)

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.

Table 3. Microphone array

Microphone	Omni-directional, Microphone Array
Characteristics	Impulse Response in Anechoic Room
Channels	8,16,64,112
Spacing	2.83cm (12kHz Sampling)
Array Design	Linear, Harmonic, 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.

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.

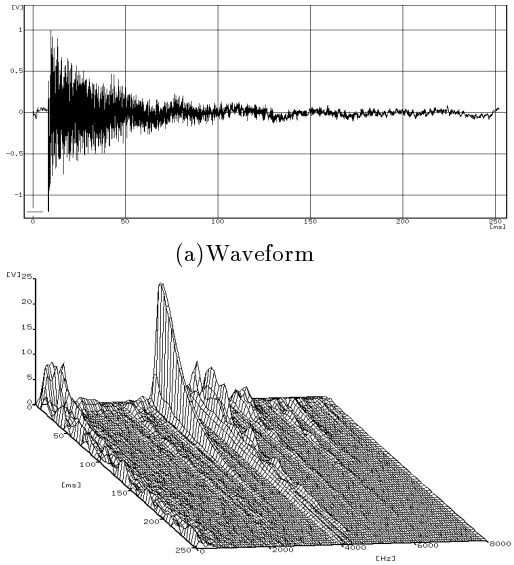
3.1.1. Specification of the data collection

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.

3.1.2. Collection of Single Sound

The sound data composed of short sounds, whose duration is about 500msec, is collected in quiet office environments for consideration of the database design.

Fifty kinds of short crash noises, whose duration is about 250msec, were collected in quiet office environments in 1997. These sounds were 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 10-20cm. The SNR to the background noise is 20-30dB. Figure 3.1.2. shows a waveform and a spectrogram of a sound signal striking a metal can by a metal stick.



(a)Waveform

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

Fig.3 Waveform and spectrogram striking a 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 7 shows the set up of the microphone array measurement. The 14ch microphone array system is used to examine the microphone array characteristics. Figure 3.2.2. shows an overview of the experiment room whose reverberation time is about 180msec. TSP signal (time stretched pulse) is used to measure the impulse responses.

3.2.2. Collected Sounds

Table 6 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.

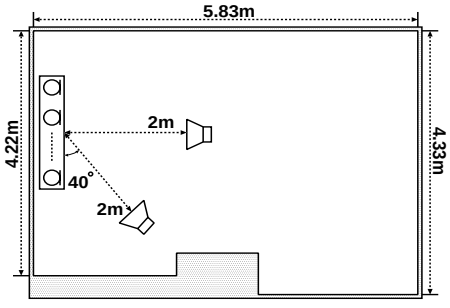


Fig.4 Sound source location

Table 6. Database using a microphone array

Data	Sound direction	Utter.	Sampling Rate	Environment
Phonetically Balanced Words	90	216	12kHz	Experiment Room
White Gaussian Noise	90/40	2	-	-
Computer Noise	90/40	2	-	-
No Source	90/40	1	-	-
Impulse Response	90/40	1	48kHz	Experiment Room
Impulse Response	90	1	-	Sound Proof Room

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 3.2.3. shows the system. The system is composed of following three components,

- A two dimensional microphone array and a three dimensional microphone array,
- 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.

Table 7. Microphone array system

Items	Specifications
Playback	Loud Speaker
M. Array	14ch, 2.83cm Linear Array
A/D,D/A	16 ch Synchronous AD/DA

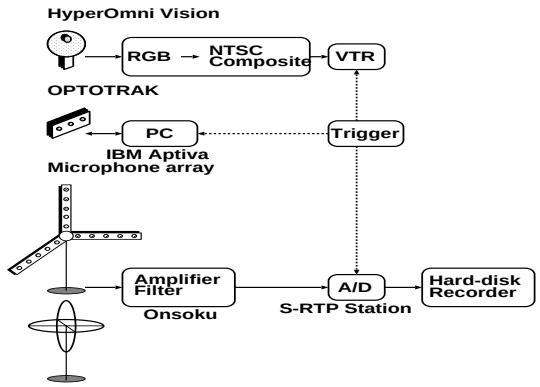


Fig.5 3-D sound scene data collection system

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. 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, P.Green, "Design,collection and analysis of a multi-simultaneous-speaker corpus", Proc.of the Institute of Acoustics, Vol.16, Part 5, pp.183-190, 1994

[2] Q.Lin,C.Che,J.French, "Description of the CAIP Speech Corpus", Proc.ICASSP94, 1994

[3] E.Jan,P.Svaizer,J.Flanagan, "A database for microphone array experimentation", Proc.Eurospeech95, 1995

[4] D.Giuliani,M.Matassoni, M.Omologo, P.Svaizer, "Use of Different Microphone Array Configurations for Hands-Free Speech Recognition in Noisy and Reverberant Environment", Proc.Eurospeech97, 1997

[5] S.Itahashi, "Recent Speech Database Projects in Japan", Proc.ICSLP90, 1990

[6] K.Takeda, Y.Sagisaka, S.Katagiri, H.Kuwabara, "A Japanese Speech Database for Various Kinds of Research Purposes" Journal of Acoustical Society of Japan, 44. 10, pp.747-754, 1988.

[7] T.Kobayashi, S.Itahashi, S.Hayamizu, T.Takezawa, "ASJ Continuous Speech Corpus for Research", Journal of Acoustical Society of Japan, 48. 12, pp.888-893, 1992.