Multi-Dimensional Data Acquisition

for Integrated Acoustic Information Research

Nobuo Kawaguchi[1,2], **Shigeki Matsubara**[1,2], **Kazuya Takeda**[2,3], **and Fumitada Itakura**[2,3]
1) Information Technology Center, Nagoya University
2) Center for Integrated Acoustic Information Research, Nagoya University
3)Graduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8601, JAPAN
kawaguti@nuie.nagoya-u.ac.jp

Abstract

The Center for Integrated Acoustic Information Research (CIAIR) at Nagoya University has been collecting various kinds of speech corpora for both of acoustic modeling and speech modeling. The corpora include multi-media data collection in moving-car environment, collection of children's voice while video gaming, room acoustics at multiple points, head related transfer functions of multiple subjects, and simultaneous interpretation of the speech between English and Japanese. This paper introduces these multi-dimensional data acquisition activities in CIAIR, and gives the basic information of the collected databases.

1. Introduction

Recently, large-scale speech corpora play an important role for both of acoustic modeling and speech modeling in the field of robust speech recognition. High-performance computer and large disk space enable to handle a huge database. The Center for Integrated Acoustic Information Research (CIAIR) at Nagoya University has been collecting various kinds of speech corpora. The focus of CIAIR is to tackle the robust speech recognition and understandings under the various environments and situations. The corpora include multi-media data collection in moving-car environment, collection of children's voice while video gaming, room acoustics at multiple points, head related transfer functions of multiple subjects, and simultaneous interpretation of the speech between English and Japanese. This paper introduces these multi-dimensional data acquisition activities in CIAIR, and gives the basic information of the collected databases.

2. In-Car Speech Database

Human-machine speech interface in a car is an important application of spoken language systems. Development of an in-car speech interface has to deal with the following two issues:

1. Noise robustness of speech
2. Continuous change of the car environment.

Towards a natural in-car speech communication environment, a large-scale corpus with multimedia data such as video images and vehicle related data is required. The in-car speech database[1] consists of (1) phonetically balanced sentences,(2)digit strings

(1) Isolated words
(2) Transcribed spoken dialogues between drivers and information systems for navigation and information retrieval.

These data are collected in vehicles under both idling and driving situations. The language of the corpus is currently Japanese. Only a few sessions are collected in English for the demonstration purpose.

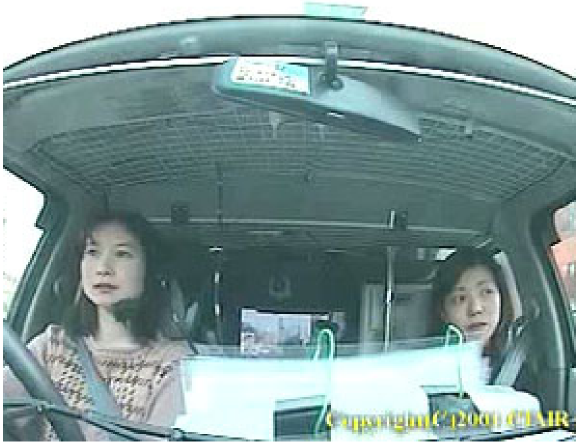


Figure 1: Dialog Recording

The number of subjects is currently about 800, total recording time is over 600 hours and total corpus size is about 2TByte. We have also been recording video images from three different angles, vehicle-control signals, and vehicle location, all synchronized with the speech recording. Figure 1 shows the example of the video image while dialog recordings.

Table 1：Collected Data

1999's collection	
Spoken dialog with human navigator	11 min
PB sent. (Idling)	50 sent.
PB sent. (Driving)	25 sent.
Isolated words	30 words.
Digit Strings	4digit*20
2000-2001's collection	
Spoken dialog with human navigator	5min
Spoken dialog with WOZ system	5min
Spoken dialog with ASR system	5min
PB sent. (Idling)	50 sent.
PB sent. (Driving)	25 sent.
Isolated words	30 words.
Digit Strings	4digit*20

Table 2: Specification of recorded data

Speech	16kHz, 16bit, 8ch
Video	MPEG-1, 29.97fps, 3ch
Control Signal	Pressure of Accelerator and Brake, Angle of Handle
Location	Differential GPS

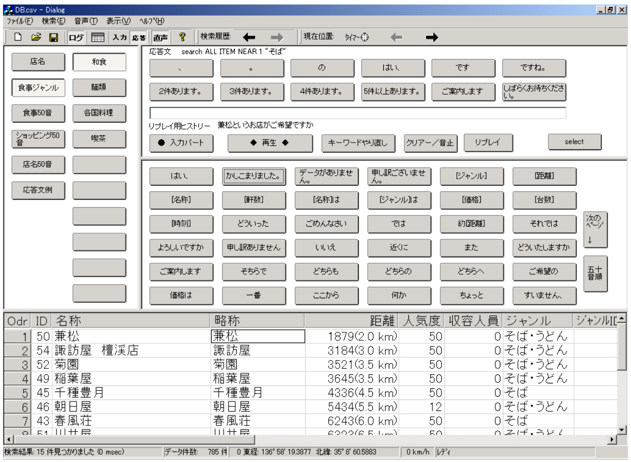
The speech data of the dialogue has been phonetically transcribed and is divided into the utterance segments that do not include pauses longer than 300 milliseconds. The speech data has been tagged with a time code. The tagging is done separately on the utterances of the driver and the operator. On the average, there are 380 utterances and 2768 morphemes in the data for a driver.

Speech data of read text has also been collected from the drivers. Each subject has read 50 phonetically balanced sentences while idling in the car and 25 sentences while driving the car. While idling, subjects use a printed text to read the phonetically balanced sentences. However, it is dangerous to read a text while driving, subjects are prompted each phonetically sentences from the head-set using special equipped wave-playback software. The speech data of the read text is mainly used for training acoustic models. Details of the collection are shown in Table 1. Table 2 shows a specification of the collected data. These multi-dimensional data are recorded synchronously, and can be synchronously analyzed.

The main concept of the dialog speech collection is to record several modes of dialogs. In 2000-2001's collection, each subject has performed a dialog with three kinds of systems. One is a human navigator, which can talk most fluently and naturally. Another is a WOZ system. Our WOZ system is equipped with a touch panel-PC and speech synthesizer. Figure 2 shows a touch screen of the WOZ system. Human navigator touches the panel while the subject makes an utterance to input the meaning of the utterance and to reply. The last system is an automatic dialog system with ASR. The system is using Julius [6] for the ASR engine. The domain of the task is the restaurant search task for all modes. Table 3 shows a basic information of the corpus. From this table, one can read the difference between three modes. "00SYS" morph/unit is 3.19 while other morph/unit are around 6.5. This means the dialog with the ASR system makes subject taciturnity.

Table 3: Basic Info. of the corpus

	99HUM	00HUM	00WOZ	00SYS
total time(sec)	141810	94692	95300	77922
sessions	209	294	293	288
speech time(sec)	97678	69390	50864	54056
driver	44559	28085	20159	11515
operator	53118	41305	30705	42541
total unit	38760	25251	19585	24944
driver	20493	12555	9831	10567
operator	18267	12696	9754	14377
total morph.	297946	215469	131569	164178
driver	137579	86567	61864	33657
operator	160367	128902	69705	130521
morph/unit	7.69	8.53	6.72	6.58
driver	6.71	6.90	6.29	3.19
operator	8.78	10.15	7.15	9.08



応答文 search ALL ITEM NEAR 1 "そば"

リプレイ用ヒストリー 兼松というお店がご希望ですか

Odr	ID	名称	略称	距離	人気度	収容人員	ジャンル	ジャンルID
1	50	兼松	兼松	1879(2.0 km)	50	0	そば・うどん	
2	54	諏訪屋　檀渓店	諏訪屋	3184(3.0 km)	50	0	そば・うどん	
3	52	菊園	菊園	3521(3.5 km)	50	0	そば・うどん	
4	49	稲葉屋	稲葉屋	3645(3.5 km)	50	0	そば・うどん	
5	45	千種豊月	千種豊月	4336(4.5 km)	50	0	そば	
6	46	朝日屋	朝日屋	5434(5.5 km)	12	0	そば・うどん	
7	43	春風荘	春風荘	6243(6.0 km)	50	0	そば	

検索結果: 15 件見つかりました (0 msec) データ件数: 785 件 0 東経: 136° 58' 19.3877 北緯: 35° 8' 60.5883 0 km/h ﾚﾃﾞｨ

Figure 2: Display of WOZ System

2.1. Data Collection Vehicle

In an ongoing project, a system specially built in a Data Collection Vehicle (DCV) has been used for synchronous recording of multi-channel audio data, multi-channel video data and the vehicle related data (Figure 3). The vehicle is equipped with eight network-connected personal computers (PCs). Three PCs have a 16-channel analog-to-digital and digital-to-analog conversion port. The data can be digitized using 16-bit resolution and sampling frequencies up to 48 kHz. One of these three PCs can be used for recording audio signals from 16 microphones. The second PC can be used for audio play back on 16 loud speakers. The third PC is used for recording signals associated with the vehicle such as the angle of the steering wheel, the status of the accelerator and brake pedals, the speed of the car, and the location information obtained from the Geographic Position System (GPS). The vehicle related control data is recorded at a sampling frequency 1kHz.

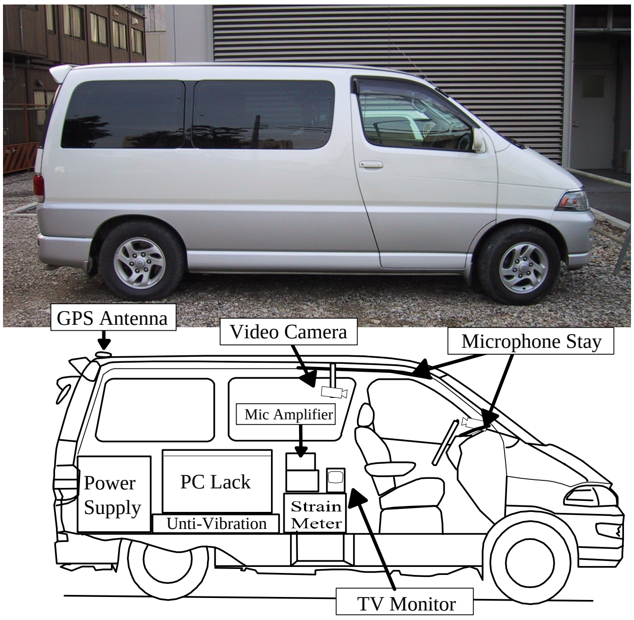


Figure 3: Data Collection Vehicle

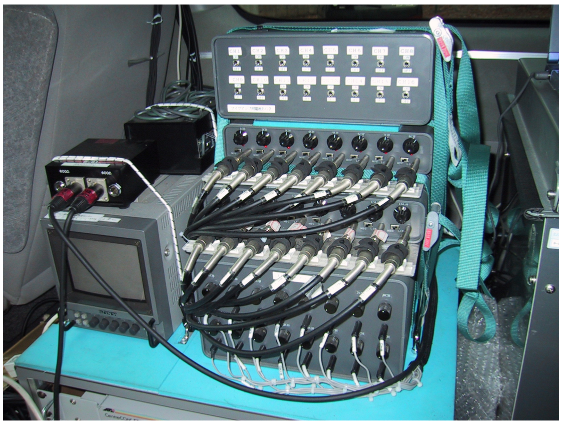


Figure 4: Microphone Amplifier

Three other PCs are used for recording video images of the driver's face, the conversational situation and the road view respectively. These images are coded into the MPEG1 format. The remaining two PCs are used for controlling the experiment. The multimedia data on all the systems is recorded synchronously. The total amount of the data is about 3 GB for about a 60-minute drive.

Figure 4 shows a 16-channel microphone amplifier and TV monitor.

3. Children's Voice Database

Robust speech recognition for young people will become more important for educational and entertainment purposes. However, it is not easy to collect a speech data especially from young children because they cannot stay long time.

To collect a speech corpus from young children, we have made special software to keep children’s attention. The software is based on a quiz game, and the answer of the quiz is the intended speech. The subjects consist of about 300 children ages from 6 to 12. Each subject has read 30 words, 30 sentences from fairy tale, and 21 command voices. The recording time is about 20 minutes for each subject.

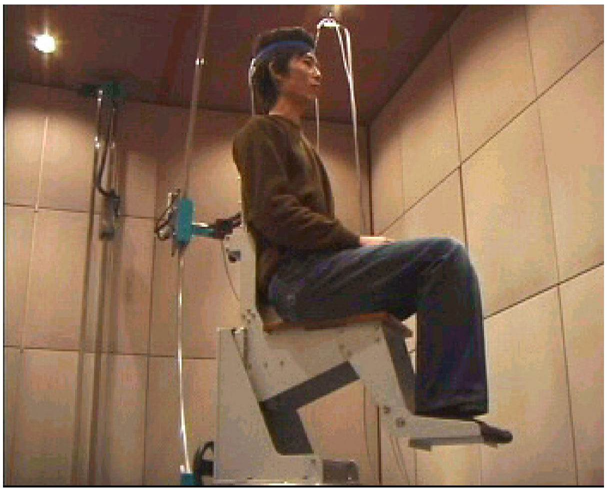


Figure 5: HRTF recording system

4. Acoustic Databases

As a multi-dimensional data acquisition, we have constructed several acoustic databases. Head related transfer functions (HRTFs) have been collected for 96 subjects. The HRTFs were measured in a reverberant room. Each HRTF is sampled in 48KHz and 512 points (10.7ms) for each 5-degree on the horizontal plane. We also made a database of room acoustics at multiple points [2]. By making full use of this database, speech recognition based on space diversity [3] might be possible. The database consists of acoustic impulse responses of 50 points. Figure 5 shows the HTRF recording system.

5. Simultaneous Interpretation

Recently, machine interpreting has become one of the important research topics with the advance of technologies for speech processing and language translation. We have made a simultaneous interpretation corpus for developing automatic simultaneous language translation system [4,5]. The corpus has the following characteristics:

(1) English and Japanese speeches are recorded in parallel.
(2) The data contains monologue speeches such as lecture and self-introduction.
(3) The exact beginning and ending times are provided for each utterance.

We have collected a total of about 70 hours of speech data and transcribed them into ASCII text files. The database consists of wave files, transcription files, and environment data files and contains about 626,000 morphemes in 66,500 utterance units.

Figure 6: Simultaneous Interpretation

6. Related Works

In this section, we describe the related works and the difference between our research.

“CU-Move” [7] is an in-vehicle speech dialogue system for route navigation and planning, which developed in Colorado University. They also perform a two-phase corpus development. First phase is for a noise analysis using several kinds of vehicles and situations, and second phase is for large-scale corpus development across the U.S. cities over 1000 subjects. Our research are quite similar, however, the differences between our research are, (1) they do not use "real" ASR system, (2) we are recording vehicle information signal such as accelerator, brake, handle and location to analyze the influence of the driving situation to the dialog, (3) we are recording multi-angle videos and distributed multi-channel microphones. They also use WOZ system but used via cellular phone. Our WOZ is onboard.

"SpeechDat-Car"[8] Project is a corpus collection project to collect data from multiple languages in an in-car setting. The effort started in April 1998 with 9 European languages. The driver is prompted to say sentences, phrases, words, letters, or numbers. However, they do not record the spoken dialog for intended task.

"VICO"[9] is a project to develop a virtual intelligent co-driver, as a robust in-vehicle spoken dialogue system. Their objective is quite similar to ours, to develop a wide-coverage basis for spoken dialogue system based on in-vehicle noise robust system.

7. Conclusion

In this paper, we presented the effort of the multi-dimensional data acquisitions for integrated acoustic information research. In-vehicle system is one of the hot-topic in the area of robust spoken dialogue system. Our corpus will play a important role in the development of in-car information system.

To construct a large-scale corpus, a lot of effort is required to make it useful. We have learned some experiences from the construction of the corpus. It may be useful for note them for other corpus development efforts.

1) Collect multi-dimensional data as much as possible for future analysis.
2) Data synchronization for multi-dimensional data must be considered at the time of system design.
3) When the data size is getting big, data transfer is heavy problem. Finally, we use the wired connection via 100Base-T.
4) Make the collection procedure automatically and flexible to adopt the various requirements.
5) Write the manual for everything.
6) Keep the record of every change of the system.

Some of our corpus is currently public, and the others will be public for free of charge after the arrangement at our WWW home page[10].

The future plans of the ongoing project for creation of the multi-dimensional data corpus include collection of multi-lingual data and collection of data in different cars.

8. References

[1] Nobuo Kawaguchi, Shigeki Matsubara, Kazuya Takeda, and Fumitada Itakura: Multimedia Data Collection of In-Car Speech Communication, Proc. of the 7th Eurpean Conference on Speech Communication and Technology(EUROSPEECH2001), pp. 2027--2030, Sep. 2001, Aalborg.

[2] Takanori Nishino, Shoji Kajita, Kazuya Takeda, and Fumitada Itakura: Interpolating Head Related Transfer Functions in the median plane, 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'99), Oct. 1999, New York.

[3] Y. Shimizu, S. Kajita, K. Takeda, F. Itakura: Speech Recognition Based on Space Diversity Using Distributed Multi-Microphone, Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2000), Jun. 2000, Istanbul.

[4] Yasuyuki Aizawa, Shigeki Matsubara, Nobuo Kawaguchi, Katsuhiko Toyama and Yasuyoshi Inagaki: Spoken Language Corpus for Machine Interpretation Research, Proc. of the 6th International Conference on Spoken Language Processing (ICSLP-2000), Vol. III, pp. 398-401, Oct. 2000, Beijing.

[5] S. Matsubara, A. Takagi, N. Kawaguchi, Y. Inagaki : Bilingual Spoken Monologue Corpus for Simultaneous Machine Interpretation Research, Proc. of LREC-2002.

[6] T.Kawahara, T.Kobayashi, K.Takeda, N.Minematsu K.Itou, M.Yamamoto, A.Yamada, T.Utsuro, K.Shikano : Japanese Dictation Toolkit: Plug-and-play Framework For Speech Recognition R\&D, Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'99), pp.393--396 (1999).

[7] J. Hansen, P. Angkititrakul, J. Plucienkowski, S.Gallant, U. Yapanel, B. Pellom, W. Ward, and R. Cole: "CU-Move": Analysis & Corpus Development for Intaractive In-Vehicle Speech Systems, Proc. of the 7th Eurpean Conference on Speech Communication and Technology(EUROSPEECH2001), pp. 2023--2026, Sep. 2001, Aalborg.

[8] P. A. Heeman, D. Cole, and A. Cronk : The U.S. SpeechDat-Car Data Collection, Proc. of the 7th Eurpean Conference on Speech Communication and Technology(EUROSPEECH2001), pp. 2031--2034, Sep. 2001, Aalborg.

[9] "VICO" Project: http://www.vico-project.org/

[10] CIAIR home page :
http://www.ciair.coe.nagoya-u.ac.jp/