CONSTRUCTION OF SPEECH CORPUS IN MOVING CAR ENVIRONMENT

Nobuo Kawaguchi[1,2] *Shigeki Matsubara*[1,3] *Hiroyuki Iwa*[1,4] *Shoji Kajita*[1,5]
Kazuya Takeda[1,2] *Fumitada Itakura*[1,5] *Yasuyoshi Inagaki*[1,2]

[1]Center for Integrated Acoustic Information Research (CIAIR), Nagoya University
[2]Graduate School of Engineering, Nagoya University
[3]Faculty of Language and Culture, Nagoya University
[4]Kojima Press Industry Co. Ltd. [5]Center for Information Media Studies, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603 Japan.
Email: kawaguti@nuie.nagoya-u.ac.jp

ABSTRACT

The Center for Integrated Acoustic Information Research (CIAIR) at Nagoya University has been collecting speech corpora in moving cars which are made available as resources to advance the research and development of robust ASRs and spoken dialogue systems under high-noise conditions. The speech corpus consists of (1) phonetically balanced sentences, (2) digit strings, (3) discrete words and (4) 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. The number of subjects is currently about 300, total recording time is over 200 hours and total corpus size is about 160GByte. We have also been recording video images from three different angles, vehicle-control signals, and vehicle location, all synchronized with the speech recording. We report the objective of the speech corpus, the recording methods and the recording vehicle developed.

1. INTRODUCTION

To develop a robust spoken dialogue system[1] which can be used for an intelligent transportation system (ITS), robust speech recognition in a high-noise environment, and understanding of a spontaneous speech dialogue are essential. A collection of speech corpora enables the further advancement of research and development of such a system[2, 3]. The Center for Integrated Acoustic Information Research (CIAIR) at Nagoya University has been collecting speech corpora in moving cars which are made available as resources to advance the research and development of robust automatic speech recognitions and spoken dialogue systems under high-noise conditions.

Figure 1: Recording Vehicle (CIAIR-SRV)

Overall Information	
number of subjects	300
total recording time	over 200 hour
total data size	over 160 GB
sampling rate	16kHz
quantization	16bit
recording channels	2–10ch

Speech data from each subject	
idling situation	
discrete words	30 words
digit strings	4 digits× 40
phonetically balanced sentences	50 sentences
driving situation	
phonetically balanced sentences	25 sent.(driver)
phonetically balanced sentences	50 sent.(passenger)
information retrieval tasks	10–20 tasks
navigation tasks	32 intersections
overall driving distance	about 20km

Table 1: Collected data in Speech Corpus

Our in-car speech corpus is composed of (1) ATR 503 phonetically balanced sentences, (2) digit strings, (3) discrete words and (4) spontaneous conversational speech using an information retrieval system. The speech is recorded in an idling situation (static) and a driving situation (dynamic) while subjects are driving the car or sitting in the passenger seat. Recorded conversations are transcribed into timed phrases with several annotation tags. Each speech is simultaneously recorded by 6 to 10 microphones distributed throughout the car, as well as headset microphones. The speech waves are digitized to 16kHz sampling frequency and 16bit quantization. Video images are also recorded from three different angles. These videos are synchronously stored in MPEG-1 format. Control signals from the vehicle, such as pressure applied to the accelerator and brake pedals, angle of the handle, RPM of the engine and speed are also synchronously recorded. Basic information of collected data is shown in Table 1.

In this paper, we report the objective of the in-car speech corpus, the speech recording vehicle (CIAIR-SRV) devel-

Figure 2: Front panel: There are wire nets for attaching microphones

Figure 3: Ceiling: There are wire nets for attaching microphones

oped , collecton methods, and the contents of the corpus.

2. OBJECTIVE OF IN-CAR SPEECH CORPUS

Objectives of the in-car speech corpus are as follows.

1. **Production of in-car phonological model.**
   To overcome the high-noise environment, we plan to produce HMMs in a real-driving environment. We also try to collect data under various conditions of, for example, the road surface, speed, weather, wind, traffic and car.
2. **Research for in-car spoken dialogue system.**
   From the transcribed spoken dialogue corpus, we are attempting to develop an example-based spoken dialogue system. The system uses the corpus as a knowledge base in order to utilize the operator's hidden knowledge included in the dialogue. We also try to investigate the timing of utterances.
3. **Basic research for in-car multimodal human interface.**
   From the information concerning the interaction between the driver and the vehicle, we can analyze the subject's ability of speech comprehension by examining factors such as the difference in the driving history and driving condition.

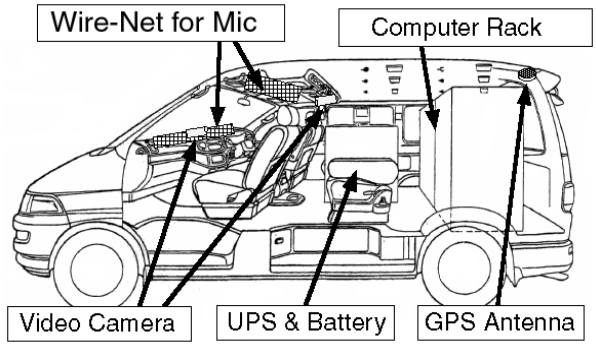

Figure 4: Configuration of CIAIR-SRV

Figure 5: Rear Seat: An operator usually sits in the rear seat

3. SPEECH RECORDING VEHICLE

We have developed a speech recording vehicle (CIAIR-SRV) in order to acquire synchronized real-time digital data (Figure 4). The vehicle contains seven network-connected computers for recording a variety of data, an extra power generator and batteries for supplying stable power, a video controller, microphone amplifiers, and speaker amplifiers.

Figures 2,3 show the wire nets to attach the microphones. There are also brackets for video cameras. Figure 5 shows a UPS and a battery case. While a subject is driving, an

Media	Spec
Sound Input	16ch,16bit,16kHz
Sound Output	16ch,16bit,16kHz
Video Input	3ch, MPEG1
Control Signal	Pressure of Accelerator and Brake, Engine RPM, Speed : 16bit,1kHz
Location	D-GPS

Table 2: Specifications of recording devices mounted in the CIAIR-SVR

Figure 6: Recording Devices: six computers and network devices are built into the vibration-free rack.

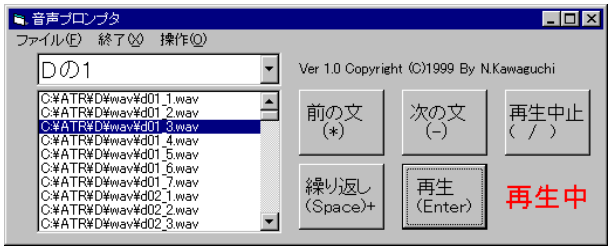

Figure 7: Speech Prompter: playback phonetically balanced sentences

operator usually sits in the rear seat. They talk to each other using headset intercoms because of in-car noise. The reason why the operator does not sit in the passenger seat is because the subject might sometimes turn his gaze to the passenger seat.

4. COLLECTION METHODS OF SPEECH CORPUS

4.1. Database for Phonological Model

We are collecting the ATR 503 phonetically balanced sentences for the production of noise-robust HMMs. Digit strings and discrete words are also collected for evaluation of the model.

In the idling situation, subjects use a printed text to read the phonetically balanced sentences. However, in the driving situation, subjects cannot read the text. Hence, we still must construct a phonological model for the driving situation. Therefore, we developed software called the "Speech Prompter" to allow the subjects to read the phonetically balanced sentences while driving the vehicle (Figure 7). Speech prompter guides the subjects by prompting phonetically balanced sentences through the headset. The timing of prompting can be controlled by the operator. We

Number of Subjects	39
Total utterances	14856
Total morphology	107966

Table 3: Number of utterances and morphology

```
    [Basic Form]                                [Pronounciation Form]
0001 00:01:543-00:10:148 M:D:I:N:
    ちょっと                [I'm a bit]     &  チョット
    小腹 <H> が             [hungry]        &  コバラ <H> ガ
    すいたんだけど <H>                      &  スイタンダケド <H>
    この                                    &  コノ
    近くに                  [near the here] &  チカクニ
    ファーストフード店て <H> [Fastfood shop] &  ファーストフードテンテ <H>
    あるのかなあ <SB>       [exists?]       &  アルノカナー <SB>
0002 00:10:683-00:13:969 F:O:I:N:
    はい                    [Yes]           &  ハイ
    マクドナルドと          [McDonald and]  &  マクドナルドト
    モスバーガーが          [MOS burger]    &  モスバーガーガ
    ございますが <SB>       [exist.]        &  ゴザイマスガ <SB>
0003 00:14:156-00:17:905 M:D:I:N:
    (F あっ)じゃ            [Oh]            &  (F アッ)ジャ
    マクドナルドの          [McDonald]      &  マクドナルドノ
    場所を                  [the place]     &  バショオ
    教えてほしいんだけど <SB> [I want to know] &  オシエテホシンダケド <SB>
0004 00:18:092-00:21:136 F:O:I:N:
    はい                    [OK]            &  ハイ
    マクドナルドは          [McDonald]      &  マクドナルドワ
    ドライブスルーされますか <H><SB>        &  ドライブスルーサレマスカ <H><SB>
                            [drive through?]
```

Figure 8: Sample transcription of spoken dialogue

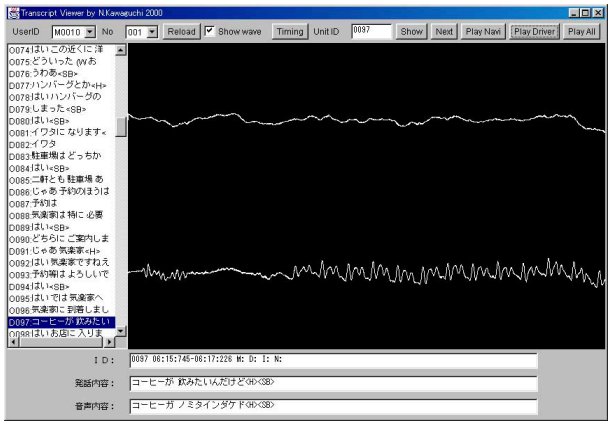

Figure 9: TransView: Tool for browsing transcription and speech signal simultaneously

divide the ATR phonetically balanced sentences to made them suitable for speech prompting.

4.2. Spoken Dialogue Corpus

Spontaneous dialogues are recorded using the Wizard of Oz method. An operator rides with a subject, and answers queries. The operator also acts as a navigator and guides the subject as to the driving route. To induce the subject to perform a information retrieval task spontaneously, we use the "prompting panel" which displays some keywords or a virtual situation.

We adopt two levels of prompting panels. A base-level panel is called the "direct panel". The direct panel displays a word, such as "Fast-food", "Bank", "Japanese-food", or "Parking". When the subject is shown the direct panel by the operator, the subject should start to ask about the place shown. In the pre-experiment, it was found that subjects sometimes just repeat the words on the panel. This makes the utterances bland.

In order to collect rich and varied utterances, we introduce the second-level panel, called the "situation panel". The situation panel contains sentences , such as "Today is an anniversary. Let's have a party.", "I'm so hungry. I need

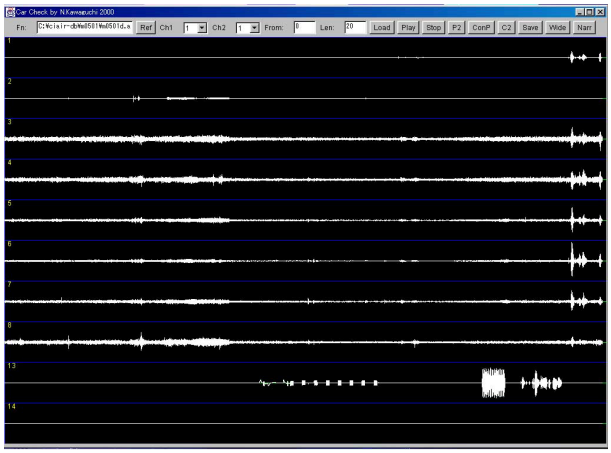

Figure 10: WaveView: Tool for browsing multi-channel speech signal

to eat!", "I'm thirsty. I want a drink!". We instructed the subjects that they should not look at the panel when they are talking. By using the situation panel, the amount of spontaneous speech dramatically increased.

5. CONTENTS OF IN-CAR SPEECH CORPUS

Figure 8 shows part of an actual transcription of our corpus (in Japanese). Each utterance is tagged with a time code and divided into several phrases. Figure 9 shows the display of the software TransView for browsing transcriptions. One can see the transcribed text and speech signal simultaneously.

Transcription is currently under way. A simple result from data analyses of 39 subjects is shown in Table 3. When the transcription of 300 subjects is complete, the total number of utterances might be 5–7 times larger than now.

Figure 10 shows the display of the software WaveView. Ten channels of the signal are simultaneously displayed. Each wave is the signal from a different microphone. Figure 11 shows the display of the software SignalView. Speed, Acceleration, Brake and Engine RPM are synchronously plotted on the screen.

Figure 12 shows a one-angle video image. MPEG-1 video is recorded duaring a dialogue-recording session.

6. CONCLUDING REMARKS

We have reported the activities for collecting a speech corpus in a moving car environment at CIAIR, Nagoya University. Our corpus will be publicly available after the appropriate arrangement. Collection of the corpus will continue for several years. After the collection and arrangement of our speech corpus, it might be recognized as a multipurpose corpus as follows.

(1) Large-scale, real-world, multichannel speech database.

(2) Multimodal database which include the information on interaction between the running vehicle and the driver.

(3) Visual database of vehicle control.

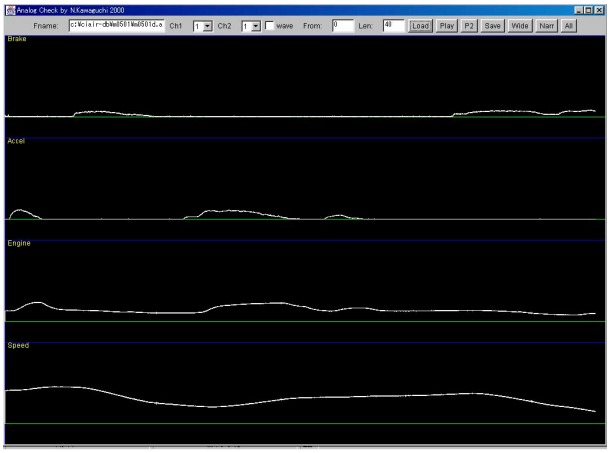

Figure 11: SignalView: vehicle control signals

Figure 12: Image from recorded video

We welcome any comments concerning our corpus, and proposals
for other data collection uses for our vehicle. More information on CIAIR corpora and distribution policies can be found at `http://www.ciair.coe.nagoya-u.ac.jp/db/` .

Acknowledgments

This research has been supported by a Grant-in-Aid for COE Research (No. 11CE2005).

7. REFERENCES

1. Nobuo Kawaguchi, Shigeki Matsubara, Katsuhiko Toyama,Yasuyoshi Inagaki: An Architecture for Multi-Domain Spoken Dialog Systems, Proceedings of the 5th Natural Language Processing Pacific Rim Symposium (NLPRS'99), pp.463–466 (1999).
2. Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, Kiyohiro Shikano and Shuichi Itahashi: JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, J. Acoust. Soc. Jpn.(E),Vol. 20, No. 3, pp.199–206 (1999).
3. Dafydd Gibbon, Roger Moore, Richard Winski (Eds.): Handbook of Standards and Resources for Spoken Language Systems, *Walter De Gruyter* (1997).