Multimedia Data Collection of In-Car Speech Communication

Nobuo Kawaguchi[1,2], *Shigeki Matsubara*[1,3],
Kazuya Takeda[1,4] *and Fumitada Itakura*[1,4]

[1]Center for Integrated Acoustic Information Research, Nagoya Univ. ,[2]Computation Center, Nagoya Univ.
[3]Faculty of Language and Culture, Nagoya Univ., [4]Graduate School of Engineering, Nagoya Univ.
Furo-cho, Chikusa-ku, Nagoya 464-8603 Japan. kawaguti@nuie.nagoya-u.ac.jp

Abstract

In this paper, we report the details of the collection of the multimedia data such as audio, video and auxiliary information of the vehicle during a spoken dialogue in a moving car. The system specially built in a Data Collection Vehicle (DCV) supports synchronous recording of multi-channel audio data from 16 microphones that can be placed in flexible positions, multi-channel video data from 3 cameras and the vehicle related data. Multimedia data has been collected for three sessions of spoken dialogue in about a 60-minute drive by each of 200 subjects. Data has been collected for two dialogue modes: (1) prompted dialogue between the driver and an accompanying operator and (2) natural dialogue between the driver and a telephone operator for information access over a cellular phone while driving a car. The corpus can be used for analysis of multimedia data in a moving car environment and also for modeling spoken dialogue in scenarios such as information access while driving a car.

1. Introduction

Human-machine speech interface in a car is an important application of spoken language systems because the conventional models of data input/output such as video display, keyboard and mouse are not convenient to use while driving a car. Development of an in-car speech interface has to deal with problems such as noise robustness and distortion of distant speech [1]. Since the background noise in a moving car is not stationary and consists of a variety of sounds, a large corpus is required for training acoustic models in the presence of different background noise conditions [2],[3]. Another important issue is that the in-car speech communication for information access by the driver has to deal with the continuously changing environment depending on the factors such as traffic condition and the distance to the destination[4]. For a system to understand the environmental condition, it may be helpful to use audio data along with other types of data such as video images of the persons involved in dialogue, images of the road in front of the car and vehicle related data such as the angle of the steering wheel, status of the accelerator and speed of the car.

Figure 1: *The exterior of Data Collection Vehicle*

In this paper, the details of the collection of the multimedia observation data of in-car speech dialogue will be presented. The main objectives of this data collection are as follows: 1) training acoustic models for the in-car speech data under various driving conditions, 2) training language models of spoken dialogue for different task domains related to information access while driving a car, and 3) modeling communication by analyzing the interaction among different types of multimedia data. 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. About 400 GB of data has been collected by recording three sessions of spoken dialogue in about a 60-minute drive by each of 200 drivers. Speech data of read text has also been collected from the drivers in the following different conditions: (1) while driving a car and (2) while idling in a car. The collected data can be used for analyzing and modeling the effects of in-car environment while driving and idling and also the effects of the natural dialogue.

In the next section we describe the multimedia data collection system in a car. In Section 3 we present the details of the data collected and the methodology used

Figure 2: *Multi-Angle Video Playback*

for data collection. A prototype spoken dialogue system for guiding the car driver about finding a restaurant is described in Section 4.

2. Data Collection Vehicle

The Data Collection Vehicle (DCV) is a car specially designed for collection of multimedia data. 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 that can be used for recording and playing back of data. 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, the rotational speed of the engine motor and the location information obtained from the Global Positioning System (GPS). The vehicle related data is recorded at a sampling frequency of 1 kHz.

Three other PCs are used for recording video images. The first camera captures the face of the driver. The second camera captures the conversation between the driver and an accompanying person acting as a navigator. The third camera captures the images of the road from the front window of the car. These images are coded into the MPEG1 format. Figure **??** shows the synchronized playback of the multi-angle videos. 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 2 GB for about a 60-minute drive during which three sessions of dialogues are recorded. The recorded data is directly stored in the hard disks of the PCs in the car.

Table 1: *Specifications of recording devices.*

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

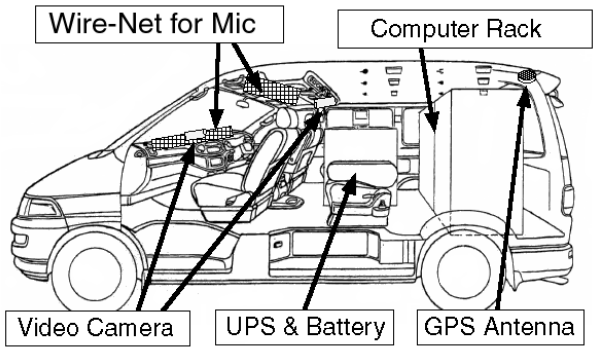

Figure 3: *Configuration of DCV*

Figure 4: *The interior of the DCV. Wire nets are attached for flexible arrangement of microphones.*

Figure 3 shows the arrangement of equipment in the DCV including the PCs, a power generator with batteries, video controller, microphone amplifiers and speaker amplifiers. An alternator and a battery are installed for stabilizing the power supply. Figure 4 shows the interior of the DCV. Wire nets are attached to the ceiling of the car so that the microphones can be arranged in flexible positions. Figure 5 shows the information viewer of the

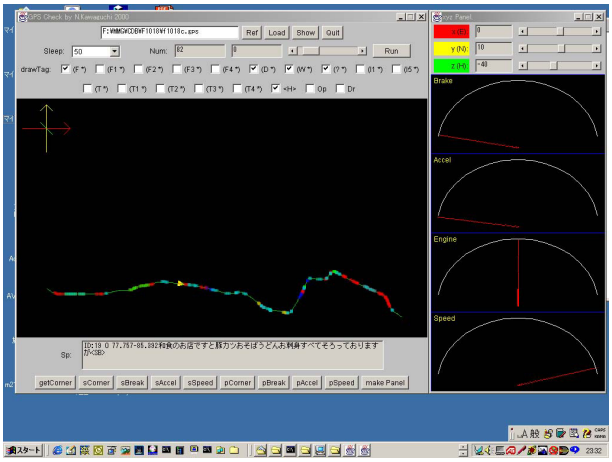

Figure 5: *GPS and Driving Information Viewer*

vehicle related data such as location information fromn GPS, the status of brake and accelerator pedals, the rotational speed of the engine motor, and the speed. By using this viewer, transcription information can be also plotted together with GPS data, one can check the dialogue context such as road straightness.

3. Speech Materials

The collecting speech materials are listed in the Table 2. Two modes of dialogue used in collection of data are as follows: (1) Prompted dialogue mode and (2) Natural dialogue mode. In the prompted dialogue mode, an operator accompanies the driver and answers the queries on the restaurants near Nagoya University campus. The operator also acts as a navigator and guides the subject about the driving route. The operator is requested to ensure that the conversation does not digress from the task domain such as information retrieval and navigation. Therefore the operator does not respond to the questions on the out-of-domain topics. In order to initiate the subject to perform an information retrieval task spontaneously, a 'prompting panel' that displays some keywords and sentences is used. Typical keywords used are as follows: Fast food, Bank, Japanese food, Parking. Typical sentences used are as follows: 'Today is an anniversary. Let's have a party.', 'I am so hungry. I need to eat!' and 'I am thirsty. I want a drink!'. In the natural dialogue mode, the driver calls the telephone operator at the NTT yellow page service, and asks for the telephone number of a particular shop using a cellular phone. The natural dialogue data has been collected while idling as well as while driving the car.

Table 2: *Speech materials recorded in the experiment.*

item	approx. time
prompted dialogue	5 min.
natural dialogue	5 min.
dialogue with system	5min.
dialogue with WOZ	5 min.
P.B. sentence (driving)	10 min.
P.B. sentence (idling)	5 min.

The speech data of the dialogue has been phonetically transcribed and is also 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 so that the timing analysis of the utterances can be carried out. 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 100 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. A speech prompter is used for collection of the data while driving. The speech data of the read text is mainly used for training acoustic models. Additionally, isolated utterances of digit strings and car control words are also collected from the drivers.

Currently, the speech data has been collected using six distant microphones and two close-talking microphones, one each for the driver and the operator.

4. Data Collection Using an ASR System

As the dialogue between man and machine is one of our final goals, we are collecting man-machine dialogues using a prototype spoken dialogue system. The task domain of the prototype system is the restaurant guidance. Drivers can get information and make a reservation of the restaurant near the campus through a conversation with the system. The automatic speech recognition module of the system is based on Julius 3.1 [8]. A bigram language model of 500-word vocabulary is trained using about 2000 sentences. Half of the training sentences are extracted from the human-human dialogue collected in the early stage of the experiment. The other sentences are generated from a finite state grammar that accepts permissible utterances in the task domain. State clustered triphone hidden Markov models consisting of 3000 states are used as acoustic models. The number of mixtures for each state is 16. The models are trained using 40,000 phonetically balanced sentences uttered by 200 speakers recorded in a soundproof room with a close-talking microphone [9]. The same microphone as in this recording is used for the speech input of the prototype dialogue system. A preliminary evaluation of the speech recognition module of the system has given a word accuracy of about

90%.

The text-to-speech module of the prototype system is based on the waveform concatenation method. The dialogue flow is modeled by the transitions among 12 dialogue states. Each state has a set of predicates on the transitions to different states. Each transition has a set of actions to be invoked. The predicates and the actions are defined in the working memory. In addition, Wizard of Oz system is also implemented on the same domain. In the experiment, the operator does not give any verbal response, but selects a keyword or a list of keywords so that the dialogue manager can appropriately change the dialogue state. The text-to-speech module gives the response to the driver.

5. Summary

In this paper we presented the details of an in-car multimedia data collection of synchronous recording of multi-channel audio data, multi-channel video data and the vehicle related data. The methodology used in collection of data for different modes of spoken dialogue by the car driver in an information access task domain is described. Data collected from 200 driver subjects can be used for analysis of the characteristics of multimedia data and developing in-car spoken dialogue systems. Modules of a prototype spoken dialogue system for restaurant information access task have also been described. The future plans of the ongoing project for creation of the in-car multimedia data corpus include collection of multilingual data and collection of data in different cars.

Acknowledgments

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

6. References

[1] J.C. Junqua and J.P. Haton: Robustness in automatic speech recognition. Kluwer Academic Publishers, 1996.

[2] P. Gelin and J.C. Junqua: Techniques for robust speech recognition in the car environment Proc. of European Conference Speech Communication and Technology (EUROSPEECH '99, Budapest 1999)

[3] M.J.Hunt: Some experiences in in-car speech recognition Proc. of the workshop on Robust Methods for Speech Recognition in Adverse Conditions (Tampere 1999) pp.25–31

[4] Deb Roy: "Grounded" Speech Communication, Proc. of the International Conference on Spoken Language Processing (ICSLP 2000, Beijin), pp.IV69–IV72 (2000)

[5] Petra Geutner, Luis Arevalo and Joerg Breuninger: VODIS - Voice-operated driver information systems: a usability study on advanced speech technologies for car environments. Proc. of International Conference on Spoken Language Processing (ICSLP2000, Beijin), pp.IV378–IV381 (2000).

[6] Asuncion Moreno, Borge Lindberg, Christoph Draxler, Gael Richard, Khalid Choukri, Jeff Allen, Stephan Eule: SpeechDat-Car: A Large Speech Database for Automotive Environments. Pro. of 2nd Int'l Conference on Language Resources and Evaluation (LREC 2000, Athens)

[7] Nobuo Kawaguchi, Shigeki Matsubara, Hiroyuki Iwa, Shoji Kajita, Kazuya Takeda, Fumitada Itakura and Yasuyoshi Inagaki: Construction of Speech Corpus in Moving Car Environment, Proc. of International Conference on Spoken Language Processing(ICSLP2000, Beijin), pp.362–365 (2000).

[8] 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)

[9] 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).