In-car Speech Data Collection along with Various Multimodal Signals

Akira Ozaki[1], Sunao Hara[1], Takashi Kusakawa[1], Chiyomi Miyajima[1], Takanori Nishino[1], Norihide Kitaoka[1], Katunobu Itou[2], Kazuya Takeda[1]

[1]Nagoya University
Furo-cho, Chikusa-ku Nagoya Aichi, 464-8603 Japan
{ozaki, hara, kusakawa, miyajima, nishino, kitaoka, takeda}@sp.m.is.nagoya-u.ac.jp
[2]Hosei University
Kajino-cho, Koganei Tokyo, 184-8584 Japan
it@fw.ipsj.or.jp

Abstract

In this paper, a large-scale real-world speech database is introduced along with other multimedia driving data. We designed a data collection vehicle equipped with various sensors to synchronously record twelve-channel speech, three-channel video, driving behavior including gas and brake pedal pressures, steering angles, and vehicle velocities, physiological signals including driver heart rate, skin conductance, and emotion-based sweating on the palms and soles, etc. These multimodal data are collected while driving on city streets and expressways under four different driving task conditions including two kinds of monologues, human-human dialog, and human-machine dialog. We investigated the response timing of drivers against navigator utterances and found that most overlapped with the preceding utterance due to the task characteristics and the features of Japanese. When comparing utterance length, speaking rate, and the filler rate of driver utterances in human-human and human-machine dialogs, we found that drivers tended to use longer and faster utterances with more fillers to talk with humans than machines.

1. Introduction

Since our modern motorized society continues to emphasize driving safety, comfort, and convenience, advanced driver assistance systems have been developed, including adaptive cruise control, lane-keeping assistance systems, and car navigation systems with speech interfaces. Future research directions will focus on developing intelligent technologies for enhancing interaction between humans and vehicles.

For such research purposes, we are constructing a large-scale real-world speech database along with other multimedia driving data. We designed a data collection vehicle equipped with various sensors to synchronously record speech, video, driving behavior, and physiological signals. Driver speech is recorded with twelve microphones distributed throughout the vehicle. Face images, a view of the road ahead, and foot movements are captured with three CCD cameras. Driving behavior signals including gas and brake pedal pressures, steering angles, vehicle velocities, and following distances are also recorded. Physiological sensors are mounted to measure driver heart rate, skin conductance, and emotion-based sweating on palms and soles to detect stress (Abut et al., 2007).

Multimodal data including speech are collected while driving on city streets and expressways under four different driving task conditions: driving while reading signs and billboards, being guided to an unfamiliar place by a human navigator on a cell phone with a hands-free device, repeating random four-character alphanumeric strings after hearing them, and interacting with a spoken dialog system to retrieve and play music.

The data statistics include gender, age, driving experience, and driving frequency. We investigate the response timing of drivers against navigator utterances and also compare utterance length, speaking rate, and filler rate of driver utterances in human-human and human-machine dialogs.

2. Speech and multimedia data collection in vehicle

2.1. Data collection vehicle

A data collection vehicle was designed for synchronously recording speech by multiple microphones with other multimedia data such as driving data. A TOYOTA Hybrid ESTIMA with 2,360 cc displacement (Figure 1) was used for data recording. Various sensors and synchronous recording systems were mounted on the machine (Figure 2). The design of the recording system is shown in Figure 3.

A potentiometer (COPAL M-22E10-050-50K) was used to measure steering angles, and pressure sensors (LPR-A-03KNS1 and LPR-R-05KNS1) were mounted on the gas and brake pedals. Vehicle velocity was measured based on the output of the JIS5601 pulse generator. Distance per 100 ms was obtained by multiplying pulse intervals and tire circumference. Digital signals were converted to analog signals by a D/A converter. Longitudinal, lateral, and vertical accelerations of the vehicle were recorded with a 3D-acceleration sensor (CXL04LP3) mounted between the driver and passenger seats, and vehicle positions were recorded with a GPS unit (PIONEER NAVICOM GPS-M1ZZ). Two kinds of distance sensors (SICK DMT-51111 and MITSUBISHI MR3685) were mounted in front of the vehicle to measure short and long ranges, respectively.

2.2. Microphones

Eleven omnidirectional condenser microphones (SONY ECM-77B) and a close-talking headset microphone were mounted on the vehicle to record driver speech. The recorded speech was amplified through YAMAHA amplifiers. The microphone positions are shown in Figure 4.

Figure 1: Data collection vehicle

Figure 2: Sensors mounted on vehicle

2.3. Other human sensing

Driver face images from the right and left front and the view of the road ahead are captured with three CCD cameras (SONY DXC-200A) at 29.4118 fps. The camera positions are shown in Figure 5. Figure 6 shows examples of video frames for cameras #1, #2, and #3.

An omni-directional camera mounted on the vehicle roof captures a 360 °view of the road.

To determine driver stress (Healey et al., 2005), physiological sensors are mounted to measure driver heart rate, emotion-based sweating on the palms and soles, and skin conductance. Driver heart rate is measured using a chest

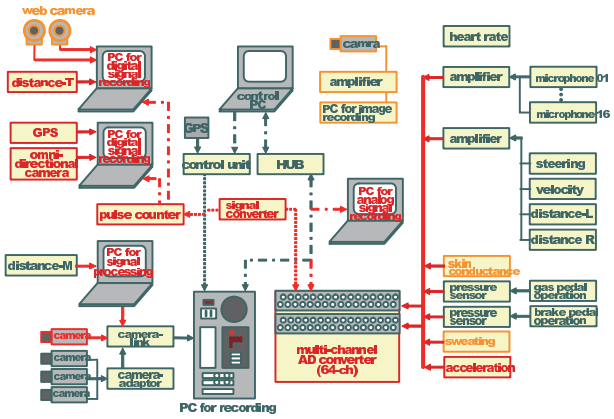


Figure 3: Block diagram of recording system

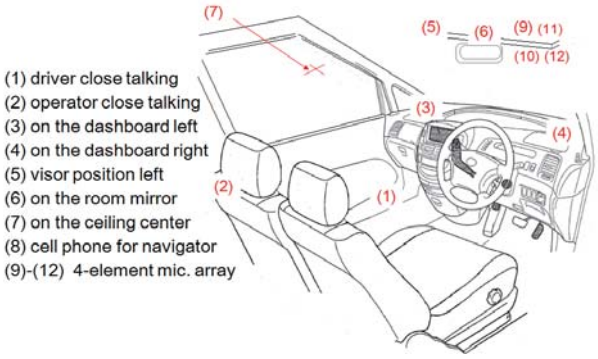


Figure 4: Microphone positions

belt sensor (POLAR S810i), sweating is measured with a perspiration meter (SKINOS SKD-2000), and skin conductance is measured with an electrodermal meter (SKINOS SK-SPA).

2.4. Synchronous recording system

For synchronous recording of the above signals, a multi-channel synchronous recording system (CORINS, MVR-303) was used. MVR-303 has a synchronous control unit and a system control PC and can record multi-channel synchronous videos and analog signals. Each PC node can store 240 GB of video data of 1.4 million pixels and 29.4118 fps that correspond to 90 minutes of video.

3. Data collection

Driving data are recorded under various conditions with four different tasks whose details are described as follows with examples of spoken sentences.

1. Signboard reading task

   Drivers read aloud such signboards as names of shops and restaurants seen from the driver seat while driving, e.g., “7-11” and “Denny’s.”

2. Navigation dialog task

   Drivers are guided to an unfamiliar place by a navigator on a cell phone with a hands-free headset. Drivers

Figure 5: Positions of video cameras

camera #1

camera #2

camera #3

Figure 6: Examples of video frames for three cameras

do not have maps, and only the navigator knows the route to the destination. The following is an example of a spoken dialog:

Navigator: You should see a restaurant on your left.
Driver: Yes, I see Kobeya.
Navigator: Well, yeah, umm, you are at the Hibarigaoka intersection. Turn left at it.
Driver: O.K.

3. Alphanumeric reading task

   Drivers repeat random four-letter strings consisting of letters of the alphabet and digits 0-9, e.g., "UKZC," "IHD3," and "BJB8." These four-letter strings are heard by earphone.

4. Music retrieval task

   Drivers retrieve and play music from 635 titles of 248 artists by a spoken dialog interface. Music can be retrieved by artist name or song title, e.g., "Beatles" or "Yesterday." An example follows:

Driver: *Ken Hirai*.
System: Do you want to search for *Ken Hirai*?
Driver: Yes, I do.
System: I found *Ken Hirai*. I'm searching for his songs...
I found three:
Grandfather's Clock, Kiss of Life, Paradise...
Driver: Good.
System: I'm playing *Paradise*.

Each driver drove for about 70 minutes. Driving data were recorded under the above four task conditions on city streets and two conditions on expressways. Driving data without any tasks were recorded as references before, between, and after the tasks. Driver resting heart rate was also recorded before and after data recording in a quiet room.
Figure 7 shows an example of the recorded data.

4. Data statistics and analysis

As of March 21, 2008, we have recorded the data of 245 subjects whose data are shown in Figure 8. Males are a little more numerous than females. The range of ages and

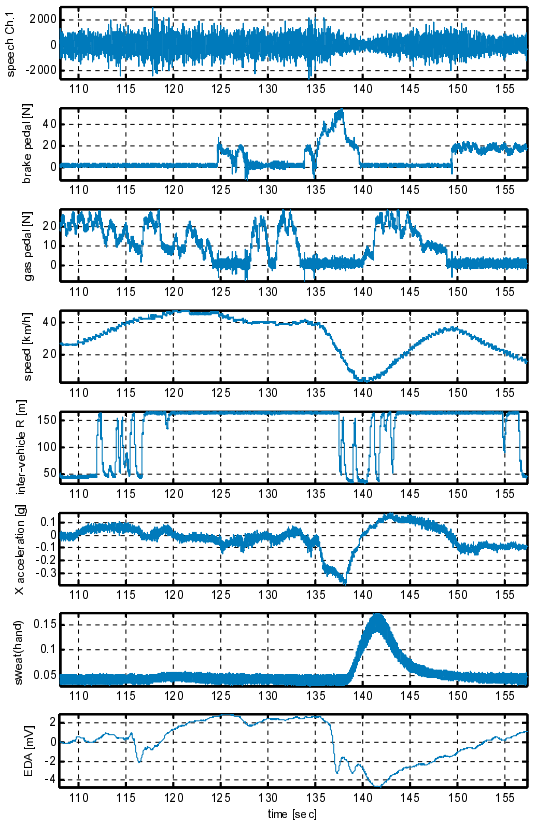
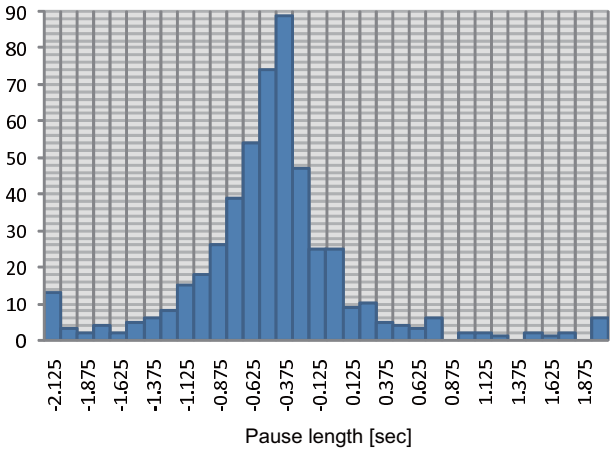
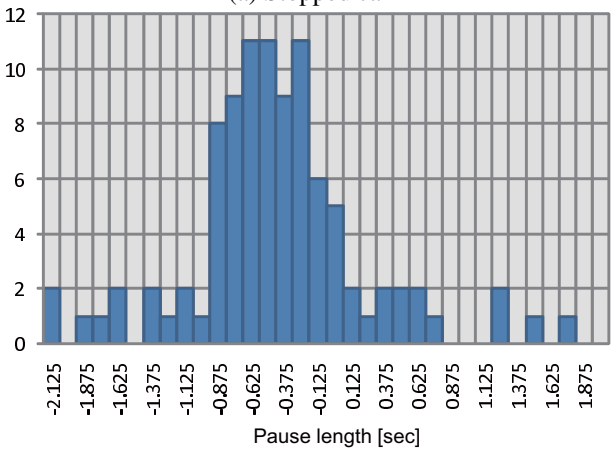


Figure 7: Examples of recorded data



(a) Stopped car



(b) Running car

Figure 9: Histograms of Pause lengths between end time of navigators' utterances and start time of succeeding drivers' utterances

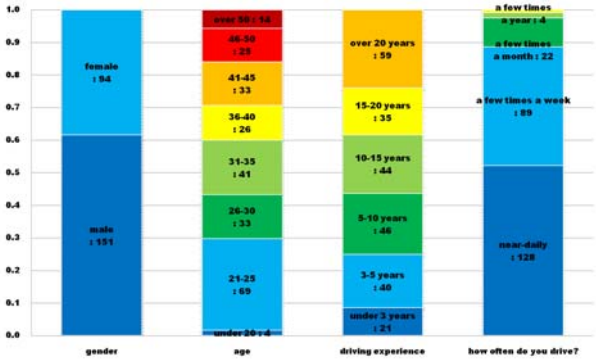


Figure 8: Statistics of subjects

driving experiences is wide and well balanced. Most of the subjects drive often in their daily lives.

We also transcribed some of their utterances (In a future, we will transcribe all utterances). The transcriptions contain utterance start and end times, gender, and transcription.

Using these data, we investigate the statistics from linguistic points of view.

4.1. Response timing of drivers

As explained in Section 3., we recorded the dialogs of drivers and navigators by cell phone. In Japan, since talking by cell phone without a hands-free microphone and speaker is prohibited, we used a hands-free headset.
Typical interactions are exampled below:

- Backchannel (which is call *Aizuchi* in Japanese) expressing approval.

  Navigator: Turn right at the next intersection.
  Driver: *Aha.*

- Repetition of navigator utterances

  Navigator: Turn left at the next intersection.
  Driver: *Turn left.*

- Status reports

  Navigator: Where are you now?
  Driver: *I can see a coffee shop called Komeda."*

From the characteristics of this task, the relation of the navigator and the driver was almost always fixed, and the former guided the latter. Here, we investigate the driver's response timing against the navigator utterances. We picked

Table 1: Statistics of speech data in navigation and music retrieval tasks. (H-H) and (H-M) indicate human-human and human-machine dialogs, respectively.

Task	Driving condition	#Utterances	#Syllables per utterance	Average utterance length (sec.)	Average speaking rate (mora/sec.)	Filler rate
Navi (H-H)	City streets	1987	8.46	1.37	6.18	0.177
Music retrieval (H-M)	City streets	698	4.06	0.697	5.83	0.072
	Expressways	878	4.08	0.675	6.05	0.072

route guidance utterances of the navigators and then calculated the pause lengths between the end time of the navigator utterances and the start time of the subsequent driver utterances:

$$\textit{Pause-length} = \textit{start-time-of-navigator's utterance} - \textit{end-time-of-driver's utterance}$$

The value less than zero means that the driver utterance overlaps the navigator utterance.

We selected twenty subjects (ten males and ten females) from the data and obtained 607 navigator-driver utterance pairs. Histograms of the pause lengths are shown in Figures 9(a) and (b), which correspond to stopped and running cars, respectively.

The average pause lengths was -0.495 [sec]. The difference between average pause lenghts for stopped and driving cars were -0.497 [sec] and -0.513 [sec], respectively, which is not significant.

We found many overlapping utterances. Drivers tended to reply quickly to confirm the contents of navigator utterances for confident guidance on unfamiliar roads. In Japanese, important keywords for route guidance tend to appear in the middle of sentences, not at the end. So the drivers often responded to the navigators after catching the keywords, resulting in responses that overlapped the preceding utterances. The latter reason is very language-specific, so we may find a language-dependent tendency when comparing the statistics of other databases recorded in other countries (Abut et al., 2007; Angkititrakul et al., 2007).

4.2. Utterance differences between human-human and human-machine dialogs

We used navigation dialogs (human-human dialogs) and music retrieval dialogs (human-machine dialogs) in the city/on the highway of 58 subjects (38 males and 20 females), all of whose utterances were already transcribed.

Since we recorded all the utterances during task completion, many out-of-task utterances, unclear pronunciations, and noise-only segments such as lip noises and sniffles are included. Handling such phenomena is indeed crucial and unavoidable, but in this paper we removed them from the data and used the rest as available data, which are shown in Table 1.

First, the average utterance lengths in syllables (#syllables per utterance) and in time (average utterance length (sec.)) are different between human-human and human-machine dialogs. This discrepancy greatly reflects the task differences, but both tasks consist of simple utterances. Human utterances tend to contain more information in human-human dialogs than in human-machine dialogs (Yamada et al., 2006), and our result is consistent with this tendency.

The average speaking rate of the human-human dialogs is slightly larger than the human-machine dialogs, but it is not so significant. Despite this evidence, the utterances in the human-human dialogs give more fluent and frank impressions. Acoustically speaking, coarticulation may occur more heavily in human-human dialogs, because we hear more ambiguous pronunciations. From a linguistic viewpoint, many more fillers appear in the human-human dialogs. Table 1 shows the filler rate, defined as utterances including filler(s)/all utterances. The existence of these phenomena indicates the difficulty of speech recognition/transcription of human-human dialogs.

Comparing driving conditions with the music navigation task, the speaking rate on city streets is lower than on expressways. The speaking rate is affected by the cognitive load (Yamada et al., 2006), and thus city streets may increase driver load.

Our data contain other signals including signals from human-sensing. Such data can be used for additional quantitative analysis of the relation between spoken dialogs and cognitive loads.

5. Conclusion

This paper summarized our on-going data collection of real-world driving at Nagoya University and highlighted some statistics of the data. Driving behavior characteristics differ from country to country based on cultural and social backgrounds. We are collaborating internationally with research groups in the U.S. and Europe to share worldwide driving data (Angkititrakul et al., 2007; Healey et al., 2005). Since similar data collection vehicles have also been developed in the U.S. and Europe, data collection is currently underway in three areas of the world. A multimodal database will be published for such research purposes as noise robust speech recognition in car environments.

6. References

H. Abut, H. Erdogan, A. Ercil, et al., "Data collection with "UYANIK": too much pain; but gains are coming," Proc. Biennial on DSP for in-Vehicle and Mobile Systems, 2007.

P. Angkititrakul and J. H. L. Hansen, "UTDrive: The smart vehicle project," Proc. Biennial on DSP for in-Vehicle and Mobile Systems, 2007.

J. A. Healey and R. W. Picard, ”Detecting stress during real world driving tasks using physiological sensors,” IEEE Trans. Intelligent Transportation Systems, vol. 6, no. 2, pp. 156-1566, 2005.

S. Yamada, T. Itoh, and K. Araki, ”Linguistic and acoustic features depending on different situations — The experiments considering speech recognition rate —,” Proc. of ICSLP-2006, pp. 3393–3396, 2006.