IEICE TRANS. INF. & SYST., VOL.E88–D, NO.3 MARCH 2005

LETTER *Special Section on Corpus-Based Speech Technologies*

CIAIR In-Car Speech Corpus —— Influence of Driving Status ——

Nobuo KAWAGUCHI[†a)], **Shigeki MATSUBARA**[†], **Kazuya TAKEDA**[†], ***and*** **Fumitada ITAKURA**[††], ***Members***

SUMMARY CIAIR, Nagoya University, has been compiling an in-car speech database since 1999. This paper discusses the basic information contained in this database and an analysis on the effects of driving status based on the database. We have developed a system called the Data Collection Vehicle (DCV), which supports synchronous recording of multi-channel audio data from 12 microphones which can be placed throughout the vehicle, multi-channel video recording from three cameras, and the collection of vehicle-related data. In the compilation process, each subject had conversations with three types of dialog system: a human, a "Wizard of Oz" system, and a spoken dialog system. Vehicle information such as speed, engine RPM, accelerator/brake-pedal pressure, and steering-wheel motion were also recorded. In this paper, we report on the effect that driving status has on phenomena specific to spoken language

key words: *speech corpus, in-car speech, ITS*

1. Introduction

The Center for Integrated Acoustic Information Research (CIAIR) has been compiling a database of in-car speech and dialog since 1999. This has been done with the goals of achieving robust speech recognition in actual usage environments and improving the level of spoken dialog [1]–[7]. At CIAIR, we have also constructed a specialized speech database recording vehicle (Fig. 1), and have recorded multi-modal information including speech and video, as well as information regarding vehicle operation and position, using more than 800 subjects. (Details on the recording methods and equipment have been given elsewhere [5].) In this paper, we report on this in-car speech database and recording vehicle, and show how this data can be used to analyze the effects of driving status on spoken language phenomena which reflect the mental focus of drivers.

The unique characteristic of this speech database is that it was compiled while subjects were actually driving the vehicle, so the dialog was recorded under different conditions than in the case of normal spoken dialog. In recordings starting from 2000, we recorded dialog using a system based on the Wizard of Oz method [6] and a spoken dialog system, as well as a human operator who played the part of a mechanical system, with each of these in-car information systems acting as the second party in the dialog.

Manuscript received July 1, 2004.
Manuscript revised September 10, 2004.

[†]The authors are with Nagoya University, Nagoya-shi, 464–8601 Japan.
[††]The author is with Meijyo University, Nagoya-shi, 468–8502 Japan.
a) E-mail: kawaguti@nagoya-u.jp

DOI: 10.1093/ietisy/e88–d.3.578

2. Construction of the In-Car Speech Database

The goal of this recording was to gather data while the subject actually drove the vehicle in a real driving environment. Table 1 is an outline of the sessions recorded for each subject. In 1999, about 11 minutes of spoken dialog with a human operator (HUM) was recorded for each subject. (These dialogs have been analyzed elsewhere [6].) From 2000 onwards, we introduced spoken dialog with a Wizard of Oz system (WOZ) and a spoken dialog system (SYS) [6] to achieve more realistic recordings. We made five-minute

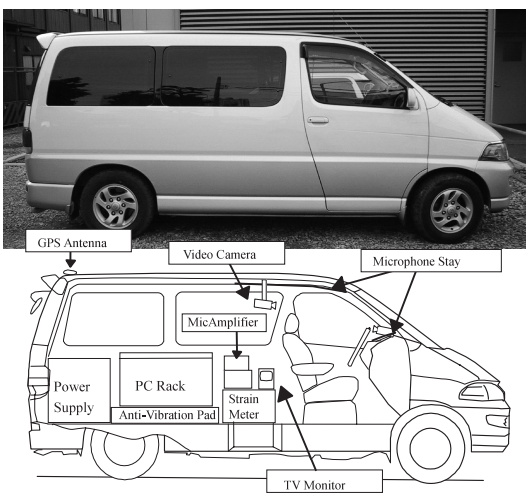

Fig. 1 CIAIR data collection vehicle.

Table 1 Collected speech data.

1999 COLLECTION	212 subj.
Spoken dialog with a human operator	11 min
PB sent. (Idling)	50 sent.
PB sent (Driving)	25 sent.
Isolated words	30 words
Digit Strings	4 digit*20
2000 - 2001 COLLECTION	**600 subj.**
Spoken dialog with a human operator	5 min
Spoken dialog with a Wizard of OZ system	5 min
Spoken dialog with a spoken dialog system	5 min
PB sent (Idling))	50 sent
PB sent (Driving)	25 sent
Isolated words	30 words
Digit Strings	4 digit*20

Table 2 Specification of collected data.

Speech	16 kHz, 16 bit, 12 ch
Video	MPEG-1, 29.97 fps, 3 ch
Vehicle Information	Status of accelerator and brake, Steering-wheel angle Engine RPM, speed: 16 bit, 1 KHz
Location	Differential-GPS (per second)

Table 3 Age/gender distribution of subjects.

Age	Male	Female	Sum
10—19	4	0	4
20—29	366	162	528
30—39	105	85	190
40—49	46	35	81
50—59	5	2	7
60 +	2	0	2
Sum	528	284	812

recordings with each system in each session. The order of dialog sessions had a significant effect. We thus used a set recording sequence for all combinations. The data shown in Table 2 was recorded during the various sessions. All dialog data was transcribed in compliance with CSJ (Corpus of Spontaneous Japanese) transcription standards [8]. After transcription, the data was divided by the type of dialog, and an "intention tag" was assigned to indicate the function of each sentence [6]. Recorded as multimedia data, the database for this test resulted in a data volume equivalent to nearly four CDs per subject. Each subject also completed questionnaires before and after the test. The questionnaire before the test consisted of 17 items, including driving experience, experience using a voice recognition system, experience using a car navigation system, and whether the subject was good at using (electronic) devices. The questionnaire after the test consisted of seven items regarding the subject's impressions of the test, degree of satisfaction using the system, and areas where the subject would like to see improvements. This large-scale questionnaire data was combined with the test data to enable various forms of analysis.

Table 3 shows the distribution of subjects by gender and age. We attempted to gather subjects that would provide an equal male/female ratio and a broad range of ages, but because the tests were done on weekdays, the subjects included a large number of students and other young males. Furthermore, since subjects needed experience driving a car, it was difficult to find older female subjects.

3. Fundamental Information for Spoken Dialog Sessions

Here, we focus on the details of the spoken dialog sessions used in the speech database. For the 812 subjects, 1,960 sessions were recorded, totalling 187.6 hours. There are a total of 1.06 million recorded morphemes, making this one of the largest dialog corpora of its kind. The calculation of morphemes does not include fillers. Tags are attached to fillers, hesitations, misspeaking, and other error elements. At the same time, utterances are divided by pauses and designated as individual utterance units. The starting and ending times of each utterance were recorded. To clarify the characteristics of in-car dialog, we transcribed recorded dialog speech data and analyzed the characteristics of the dialog data based on the transcription. Specifically, we surveyed the drivers' utterances, focusing on utterance speed and length, as well as phenomena specific to the spoken language. Morpheme analysis was done using the Chasen [13] morpheme analysis tool.

3.1 Utterance Speed

We measured the drivers' utterance speed based on utterance time information from the transcribed text. The average number of mora/s throughout the 1,960 sessions was 5.97, which is low compared to regular dialog speech (8.5–12.5 mora/s) [9], [10] and "lecture speech" (6.5–10.5 mora/s) [11], indicating that the utterances were comparatively slow. This was probably because the drivers were concentrating on the task of driving and unable to pay full attention to the utterance task. The average utterance speed of the drivers for each dialog session was 6.01 mora/s (HUM), 6.07 mora/s (WOZ), and 5.59 mora/s (SYS). We also confirmed that the utterance speed was particularly slow during the SYS sessions. This was because drivers tend to slowly repeat themselves in response to erroneous operations such as speech recognition errors, and because of the short length of each utterance unit. The operator's utterance speed in the HUM sessions was 6.97 mora/s. Like those of the drivers, the operators' utterances were slow compared to standard dialog, because of the need to communicate information fully to the drivers. The utterance speed for the synthesized voice was 6.07 mora/s (WOZ), and 6.38 mora/s (SYS).

3.2 Utterance Unit

Upon transcription, each utterance was divided into utterance units using pauses. The number of morphemes for each utterance unit differed significantly depending on the dialog session, averaging 5.72 for HUM, 5.27 for WOZ, and 2.48 for SYS. This can be considered a major indicator of the features of each dialog session. It became clear that in the case of HUM, the utterance units contained many morphemes, but in the case of SYS the utterances were far shorter. We presume that this was because the subjects were influenced by the performance of the dialog system used for the recording, and carried out dialogs with a focus on short utterances. In the case of WOZ, where responses created by the operator were output using a synthesized voice, the average utterance unit length was shorter than, but still close to, the average utterance unit length in the case of HUM. This indicates that the low utterance-understanding capa-

Table 4 Occurrences of spoken language phenomena.

	Driver		Operator	
	Occ.	Rate.	Occ.	Rate.
Filler	31153	31.2%	20257	29.0%
Hesitation	8475	6.0%	2065	4.6%
Misspeaking	4256	2.8%	1621	3.6%

Table 5 Analysis of driving conditions.

	Utterance Time		Utterance Unit		Morphs	Mora
	s	%	units	%	/unit	/s
Driving	78363	62.1%	36833	62.9%	5.17	5.87
Idling	47883	37.9%	21723	37.1%	5.44	5.97
Sum	126346		58556		5.27	5.91

Table 6 Driving conditions and fillers.

	Filler Occ.	Filler Rate
Driving	12952	35.16%
Idling	7759	35.72%
Sum	20711	35.73%

Table 7 Analysis of pedal operation.

		Utt. Time		Utt. Unit.		morph	mora.
		s	%	units	%	/ unit	/ s
Acc. Pedal	operat.	38594	30.6	17696	30.2	5.30	5.86
	non-op	87567	69.4	40860	69.8	5.26	5.94
Brake Pedal	operat.	59923	47.5	27757	47.4	5.29	5.91
	non-op	66314	52.6	30799	52.6	5.25	5.92
Sum/Average		126161		58556		5.27	5.91

bility in the case of SYS interferes with dialog smoothness more strongly than the synthesized speech.

3.3 Phenomena Specific to Spoken Language

In terms of phenomena specific to spoken language, we focused on fillers, hesitations, and misspeakings, and examined their frequency and types. Table 4 shows the total number of occurrences and the number of occurrences per utterance unit (the appearance rate) for utterances by the drivers and the operators. Fillers occurred in 31,135 of the drivers' 84,948 utterance units; i.e., an occurrence rate of 31.2%. This figure is low compared to the results of past surveys of human dialogs [12]. This is believed to be because, as a rule, utterance units are shorter during in-car dialogs than during regular dialogs. The types of fillers that occurred in the drivers' utterances, and the order of frequencies are also similar to those found in previous studies. Similar trends were also observed regarding hesitations and misspeakings.

4. Analysis Using Vehicle Information

The main feature of this speech database is that in addition to speech data, vehicle information has been recorded at the same time. In this section, we will conduct an analysis using information related to the vehicle speed and accelerator, brake, and steering-wheel operation. Vehicle data was collected for only 1,741 of the 1,960 sessions, though, because of recording problems.

For this paper, we have analyzed only HUM and WOZ sessions. In SYS sessions, the utterance time is significantly shorter than in other sessions. In addition, the dialog style in a SYS session is completely different. Most of the dialog tends to be composed of short sentences and repeated utterances frequently occur because of speech recognition failures.

4.1 Analysis Regarding Driving Conditions

Regarding driving information, we recorded speed and engine RPM from the vehicle using pulse signals. We extracted the speed information from the vehicle information, and categorized cases where utterances overlapped with periods where the speed was 6 km/h or above for five seconds or more as "while driving" and any other period as "while idling" (Table 5). Of the total, 62%–63% of the data fell into the "while driving" category. The average number of moras was roughly the same for "while driving" and "while idling", and we found that the number of morphemes per utterance unit tended to be slightly shorter (by about 5%) while driving. A shorter average utterance unit suggests that the utterances themselves were comparatively simple. We can presume, then, that a driver must pay more attention to driving while the car is moving than when stopped, resulting in a lower degree of concentration on the dialog.

Table 6 shows the results of a survey of driving conditions and filler occurrence rates. Unlike the previous survey, we can see here that there is no significant correlation between driving conditions and filler appearance rates. (Note: The large differences in values compared to those in Sect. 3 are due to the omission of data for SYS sessions.)

4.2 Analysis Regarding Accelerator and Brake Pedal Operation

To measure accelerator and brake operation, we used pressure sensors to record the pressure generated when a driver stepped on either pedal. In this analysis, we defined a pressure of 0.5 kg on the pedal for 0.5 seconds or more as "operation", and any other time as "non-operation". The results of this survey are shown in Table 7. We found that about 30% of the utterance time coincided with accelerator operation, and about 47% with brake operation. This means that about 78% of all utterances by subjects were made while

Table 8 Analysis of filler and pedal operation.

		Filler Occ.	Filler Rate
Accelerator Pedal	operation	6421	36.29%
	non-op	14290	34.97%
Brake Pedal	operation	9840	35.45%
	non-op	10871	35.30%
Sum / Average		20711	35.37%

Table 9 Analysis of steering-wheel operation.

	Utterance Time		Utterance Unit		mora
	s	%	unit	%	/s
Straight	46749	57.1	26180	68.4	6.04
Adjustmen	33320	40.7	11582	30.3	5.81
Turning	1824	2.2	507	1.3	5.55
Sum/Ave	81893		38269		5.94

Table 10 Analysis of filler and steering-wheel operation.

	Filler Occ.	Filler Rate
Straight	8353	31.91%
Adjustment	5056	43.65%
Turning	263	51.87%
Sum/Ave	13672	35.73%

operating the vehicle using one of these pedals. Accelerator operation had a slightly greater effect than brake operation on utterance speed and utterance unit length. We can see that compared to the shorter utterance unit length during driving, accelerator operation had a smaller effect on the utterance unit length, but the effects on utterance speed were greater during accelerator operation.

Next, we examined the correlation between filler occurrence rates and pedal operation. The results (Table 8) indicate that the filler occurrence rate was slightly higher during accelerator operation than during brake operation. From this data as well, we can see that accelerator operation places a greater burden on the driver. Driving conditions affect the utterance unit length, but not filler occurrence rates; pedal operation conditions have a larger effect on filler occurrence rates than on utterance unit length.

4.3 Analysis Regarding Steering-Wheel Operation

With regard to steering-wheel operation, we applied variable resistance using a gear attached to the steering wheel to obtain data on the steering wheel position. Because steering-wheel operation information has been recorded only since 2000, here we examine HUM and WOZ sessions recorded from 2000 onward. We defined steering-wheel rotation of 15° or less as "driving straight," rotation of 15° to 180° as "adjustment," and of 180° or more as "turning." Lane changes and similar operations are examples of "adjustment."

As shown in Table 9, about 57% of utterances corresponded with "driving straight", about 40% with "adjustments", and about 2.2% with "turning". This data indicates that steering-wheel operation has a dramatic effect on the driver's utterances. As in the above analyses, utterance speed dropped by about 6%, particularly in situations such as steering-wheel operation that require the driver's attention. Conversely, when the driver was driving straight, the utterance speed increased, perhaps because the driver was more relaxed.

Table 10 shows the correlation between steering-wheel operation and the filler appearance rate. Just as utterance speed was dramatically affected by steering-wheel operation, we can see that the filler occurrence rate changed greatly depending on steering-wheel operation conditions. Particularly during steering wheel operation, fillers occurred in about half of all utterances, indicating that the driver's attention was drawn away from utterances by driving operations.

4.4 Summary of Analyses Based on Vehicle Information

We have examined the relationships between utterances and vehicle information. From the above, we have confirmed that driver utterances are influenced by driving conditions and effects are particularly noticeable with respect to accelerator and steering-wheel operation. When constructing in-car information systems based on telematics and related technologies, mechanisms should be incorporated so that it prevents information from being presented to the driver at inopportune times by utilizing information on the status of accelerator and steering-wheel operation.

5. Conclusion

In this paper, we have discussed an in-car speech database compiled by CIAIR at Nagoya University using a total of 812 subjects, and comprising 1,960 sessions totalling 187.5 hours. We studied a dialog corpus containing a total of 1.06 million recorded morphemes to determine fundamental information and phenomena specific to the spoken language. We also used vehicle information, one of the unique features of this database, to study the ways in which utterances are affected by driving conditions, accelerator and brake operation, and steering-wheel operation. Based on this study, we learned the following.

- Utterance speed for drivers, at 5.5–6.1 mora/s, was slower than that during regular dialog.
- Dialogs with systems were characterized by shorter utterance units than were used in dialogs with humans.
- Driving conditions ("while driving" vs. "while stopped") had a limited effect on dialog.
- Vehicle operation, and particularly accelerator and

steering-wheel operation, significantly affected utterances.

This database is based on a large number of subjects and features a huge volume of multimedia data, and the current study covers only a segment of this data. We look forward to seeing even more extensive research related to spoken dialog systems in real environments undertaken based on further applications of the database.

Acknowledgments

This research was conducted with the aid of a Ministry of Education, Culture, Sports, Science and Technology Grant-in-Aid for COE (Center of Excellence) Scientific Research (No. 11CE2005).

References

[1] N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura, and Y. Inagaki, "Construction speech corpus in moving car environment," Proc. ICSLP-2000, III, pp.362–365, 2000.

[2] N. Kawaguchi, S. Matsubara, K. Takeda, and F. Itakura, "Multimedia data collection of in-car speech communication," EUROSPEECH2001, pp.2027–2030, 2001.

[3] N. Kawaguchi, S. Matsubara, K. Takeda, and F. Itakura, "Multi-dimensional data acquisition for integrated acoustic information research," The 3rd International Conference on Language Resources and Evaluation, LREC-2002, vol.I, pp.2043–2046, 2002.

[4] N. Kawaguchi, K. Takeda, S. Matsubara, I. Yokoo, T. Ito, K. Tatara, T. Shinde, and F. Itakura, "CIAIR speech corpus for real world speech recognition," International Joint Conference of the 5th Symposium on Natural Language Processing, SNLPoCOCOSDA-2002, pp.288–295, 2002.

[5] H. Murao, N. Kawaguchi, S. Matsubara, Y. Yamaguchi, and Y. Inagaki, "Example-based spoken dialogue system using WOZ system log," SIGdial Workshop on Discourse and Dialogue (SIGDIAL2003), pp.140–148, 2003.

[6] I. Kishida, Y. Irie, Y. Yamaguchi, S. Matsubara, N. Kawaguchi, and Y. Inagaki, "Construction of an advanced in-car spoken dialogue corpus and its characteristic analysis," EUROSPEECH2003, pp.1581–1584, 2003.

[7] N. Kawaguchi, K. Takeda, and F. Itakura, "Multimedia corpus of in-car speech communication," J. VLSI Signal Processing, vol.36, no.2, pp.153–159, 2004.

[8] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," LREC-2000, pp.947–952, 2000.

[9] K. Hirose and M. Sakata, "Comparison on prosody of dialogue speech with read speech," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J79-D-II, no.12, pp.2154–2162, Dec. 1996.

[10] N. Murakami and S. Sagayama, "A study on acoustic characteristics on spontaneous speech," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J78-D-II, no.12, pp.1741–1749, Dec. 1995.

[11] N. Minematsu, Y. Kataoka, and S. Nakagawa, "Analysis of spoken language in lecture style," IPSJ SIG Technical Report, SIG-SLP-8, pp.39–46, 1995.

[12] S. Nakagawa and S. Kobayashi, "Acoustic characteristics and occurrence pattern of pause/filler in natural spoken dialog," J. ASJ, vol.51, no.3, pp.202–210, 2003.

[13] M. Asahara and Y. Matsumoto, "Extended models and tools for high-performance part-of-speech tagger," Proc. 17th Conf. on Computational Linguistics (COLING2000), pp.21–27, 2000.