AN ADVANCED JAPANESE SPEECH CORPUS FOR IN-CAR SPOKEN DIALOGUE RESEARCH

Yuki Irie[1], *Nobuo Kawaguchi*[1,2,3], *Shigeki Matsubara*[2,3], *Itsuki Kishida*[1], *Yukiko Yamaguchi*[2],
Kazuya Takeda[1,3], *Fumitada Itakura*[3,4,] *and Yasuyoshi Inagaki*[5]

1) Graduate School of Information Science, Nagoya University,
2) Information Technology Center, Nagoya University,
3) Center for Integrated Acoustic Information Research, Nagoya University,
4) Graduate School of Engineering, Nagoya University,
5) Faculty of Information Science and Technology, Aichi Prefectural University
Furo-cho, Chikusa-ku, Nagoya, 464-8603 Japan

yuki-i@inagaki.nuie.nagoya-u.ac.jp

Abstract

In this paper, we report the construction of an advanced in-car speech dialogue corpus and the result of the preliminary analysis. We have developed the system, specially built in a Data Collection Vehicle (DCV), which supports synchronous recording of multi-channel audio data from 16 microphones that can be placed in flexible positions, the multi-channel video data from 3 cameras and the vehicle-related data. The multimedia data has been collected for three sessions of spoken dialogue with different types of the navigator in an about 60-minute drive by each of 800 subjects. We have defined an organization of intention tags called the Layered Intention Tag and provided for each speech unit for the purpose of the analysis of dialogue structures. Then we have marked the tags to over 35,000 speech units. We have developed the dialogue sequence viewer to analyze the basic dialogue strategy of the human-navigator conversation. We also report the preliminary analysis on the relation between the intention and the linguistic phenomenon.

1 INTRODUCTION

Speech interface which can deal with spontaneous speech is one of the landmarks for human-machine interface. To attain the landmark, large-scale speech corpora play important roles for both of acoustic modeling and language modeling in the field of robust and natural speech interface. The Center for Integrated Acoustic Information Research (CIAIR) at Nagoya University has been collecting a large scale corpus of the in-car speech [1,5,6]. In-car speech interface has to deal with the dynamic situation of the driver such as traffic condition and the distance to the destination [2,8,9].

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 the acoustic models for the in-car speech data, 2) training the language models of spoken dialogues with the task domains related to information access while driving a car, and 3) modeling the communication by analyzing the interaction among the different types of the multimedia data. In our ongoing project, a system specially built in a Data Collection Vehicle (DCV)(Fig. 1) has been used for synchronous recording of multi-channel audio data, multi-channel video data and the vehicle related data. About 1.4 TB of the data has been collected by recording several sessions of spoken dialogue in an about 60-minute drive by each of 800 drivers. All of the spoken dialogues are transcribed into the texts with detailed information. We have defined the Layered Intention Tag for analyzing the dialogue structures. The data can be used for analyzing and modeling the interaction between the navigators and drivers in an in-car environment while driving and idling.

In the next section, we briefly describe the multimedia data collection in the car. In Section 3, we introduce the Layered Intention Tag for analysis of dialogue acts. The preliminary analysis on the relation between the intention and the linguistic phenomenon is presented in Section 4.

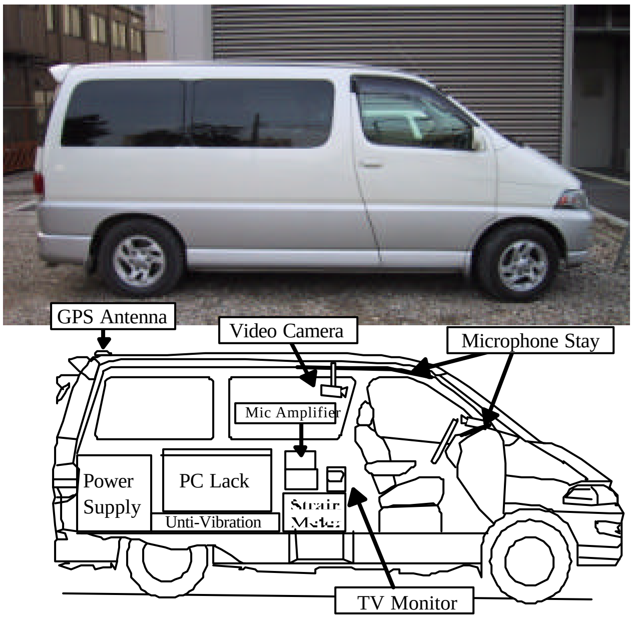

Figure 1: Data Collection Vehicle

2 IN-CAR SPEECH DATA COLLECTION

The main concept of the dialogue speech collection is to record several modes of dialogues. In 2000-2001's collection, each subject has performed a dialogue with three kinds of navigators. One is a human navigator, which can talk most fluently and naturally with the driver. Another is a WOZ system. Our WOZ system is equipped with a touch panel-PC and a speech synthesizer. Figure 2 shows a recording situation of the WOZ system. Then, a human operator 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 dialogue system with ASR. The navigator uses Julius [3] for the ASR engine. The domain of the task is the information retrieval task for all modes. Table 1,2,3 shows a basic information of the collected corpus. Please refer [6,10] for the detailed information about the corpus.

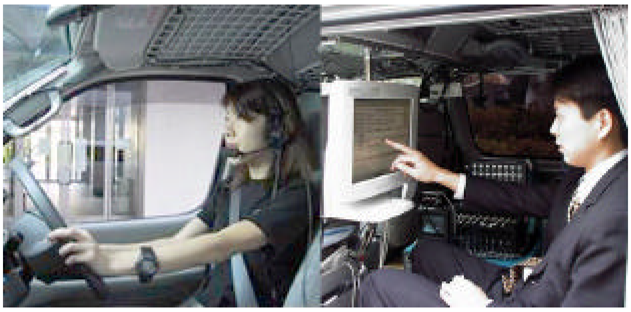

Figure 2: WOZ Dialogue Recording

Table 1: Collected Speech Data

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

Table 2: Statistics of the Corpus

	99HUM	00-1HUM	00-1WOZ	00-1ASR	Total
Rec. time(sec)	141,822	188,157	189,162	156,091	187.6 hour
Sessions	209	589	587	575	1960
Speech len.(sec)	98,100	137,025	98,288	102,933	121.2 hour
driver	44,772	54,140	38,286	22,516	44.4 hour
operator	53,328	82,885	60,002	80,417	76.8 hour
Speech unit	38,760	49,429	39,578	47,848	175,615
driver	20,493	24,540	19,076	21,289	85,398
operator	18,267	24,889	20,502	26,559	90,217

Table 3: Specification of recorded data

Speech	16kHz, 16bit, 16ch
Video	MPEG-1, 29.97fps, 3ch
Control Signal	Status of Accelerator and Brake, Angle of Steering wheel Engine RPM, Speed: 16bit 1kHz
Location	Differential GPS (each 1sec)

3 LAYERED INTENTION TAG

To develop a spoken dialogue system utilizing a speech corpus[4], we require some specified information for each sentence, which corresponds to the system reaction. Additionally, to perform the reaction to the user, we need to predict the intention of the user's utterances. By the preliminary experience, we learned that the user's intention is widely spread even in a simple task. So, if we define the detailed intention tags, we need to define dozens of them. Therefore, we divide the intention tags into several layers to simplify it. This also benefits the hierarchical analysis of the intentions.

We define the Layered Intention Tag (LIT) as shown in Table 4. LIT is composed from 4 layers. Discourse Act layer denotes the role of the speech unit in the dialogue. All of Discourse Act tags are "task independent tags". Action layer denotes the action of the speech unit. Action tags are divided into "task-independent tags" and "task-dependent tags". "Confirm" and "Exhibit" are task-independent, but the others ("Search", "ReSearch", "Guide", "Select" and "Reserve") are task-dependent tags. Object layer denotes the object of the action such as "Shop", "Parking", etc. Argument layer denotes the other miscellaneous information about the speech unit. Most of the argument layer tags can be decided directly from the specific keywords in the sentence. As Figure 3 shows, the lower layered intention tag depends on the upper layered one.

An example of a dialogue between a human navigator and a subject is shown in Figure 3. For each utterance (speech unit), we provided the intention tag. At this time, we have tagged for over 35,000 speech units. Table 5 shows the current size of our corpus with the layered tags. Table 6 shows the top 10 kinds of layered intention tags and their appearance frequencies in the corpus.

Table 4: Layered Intention Tag (a part of)

Discourse Act	Action	Object	Argument
Request(Req)	Confirm(Conf)	Shop	ShopName
Propose(Prop)	Exhibit(Exhb)	Parking	Genre
Express(Expr)	Search(Srch)	ShopInfo	Price
Suggest(Sugg)	ReSearch(ReSe)	ParkingInfo	Place
Statement(Stat)	Guide(Guid)	SearchResult	Date
	Select(Sel)	RequestDetail	Menu
	Reserve(Res)	SelectionDetail	Count
		YesOrNo	Time

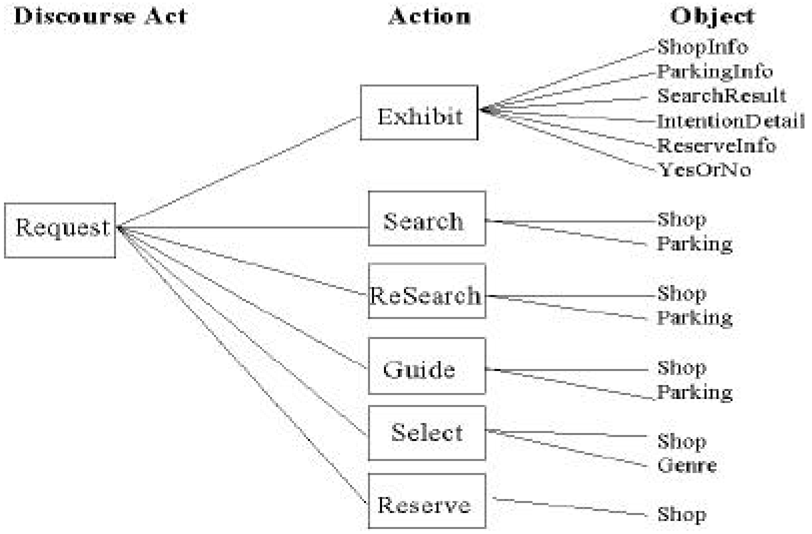

Figure 3: A part of intention structure tree

Table 5: Statistics of the intention tagged corpus

	99HUM	00HUM	00WOZ	01HUM	01WOZ
Sessions	72	297	297	295	295
Task (restaurant)	425	793	890	626	907
Speech unit	4,909	8,133	8,420	5,628	8,331
driver	2,331	3,806	3,760	2,624	3,713
operator	2,578	4,327	4,660	3,004	4,618

Table 6: Appearance frequencies of the intention tags (top 10)

	99	00	00	01	01	
	HUM	HUM	WOZ	HUM	WOZ	Total
Stat+Exhb+IntDetail	694	1,192	1,442	818	1,549	5,695
Stat+Exhb+SearchResult	665	1,303	1,260	938	1,285	5,451
Req+Srch+Shop	497	811	845	894	910	3,957
Expr+Guid+Shop	353	709	830	568	834	3,294
Stat+Sel+Shop	365	685	749	563	793	3,155
Stat+Exhb+ShopInfo	733	540	362	336	337	2,308
Req+Exhb+ShopInfo	655	377	223	259	338	1,852
Stat+Sel+Gerne	46	378	425	325	466	1,640
Req+Sel+Gerne	58	219	379	283	428	1,367
Req+ReSe+Shop	162	345	205	260	310	1,282

4 ANALYSIS OF THE CORPUS

We divide the recording session into sevral short tasks. Each task is a dialogue about a single theme. The dialogue in Figure 4 is an example of a single task about the restaurant query. We have provided the tags for the all tasks about the restaurant query. Total number of the tagged task is 3641. A task consists of 9.7 speech units on average.

```
        Utterance                         |     Intention Tag
------------------------------------------------------------------
Subj:Umm, I'm looking for
     a fastfood restaurant.                    Req+Srch+Shop

Navi:Well, there are McDonald's,
     Mr.Donuts,  and Lotteria near here.       Stat+Exhb+SrchRes

Subj:So, McDonald's please.                    Stat+Sel+Shop

Navi:OK. I'll navigate to the
     McDonald's restaurant.                    Expr+Guid+Shop
```

Figure 4: Example of the transcription with LIT

4.1 Dialogue Sequence Viewer

To understand and analyze the dialogue intuitively, we have developed a dialogue sequence viewer shown in Figure 5. We combine the units into a 'turn' which means a change of a speaker. So, each turn may have several tags. Each node means a tag with a turn number, and link between nodes means a sequence of the dialogue. The thickness of a link means an occurrence count of the tag's connection. Figure 5 only shows the case of short dialogues which ends only 4 turns. The average turn count of the restaurant query task is about 10. By using the dialogue viewer, we found that most of the dialogue sequences pass through the typical tags such as "Req+Srch+Shop", "Stat+Exhb+SrchRes", "Stat+Sel+Shop", and "Expr+Guid+Shop". The dialogue in Figure 4 is one of the typical sequences. We also check the dialogue of the length 6, 8 and 10. From this experience, we notice that the start section and the end section of the dialogue are very similar in those of the dialogue, the length of dialogues which is different each other.

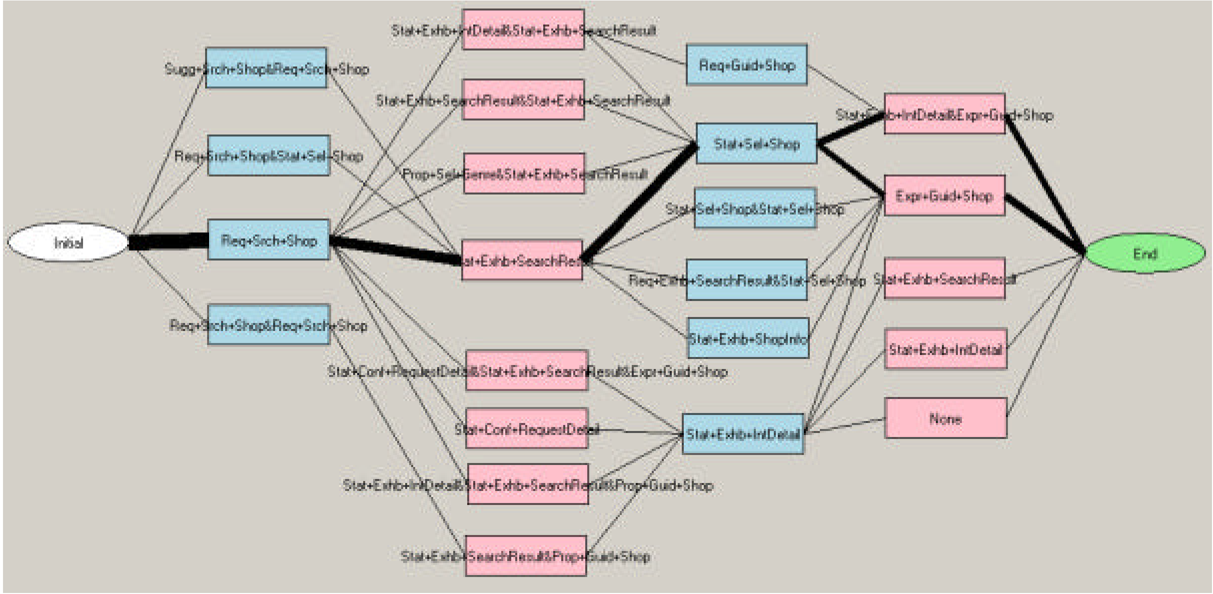
Initial
Sugg+Srch+Shop&Req+Srch+Shop
Req+Srch+Shop&Stat+Sel+Shop
Req+Srch+Shop
Req+Srch+Shop&Req+Srch+Shop
Stat+Exhb+IntDetail&Stat+Exhb+SearchResult
Stat+Exhb+SearchResult&Stat+Exhb+SearchResult
Prop+Sel+Genre&Stat+Exhb+SearchResult
Stat+Exhb+SearchResult
Stat+Conf+RequestDetail&Stat+Exhb+SearchResult&Expr+Guid+Shop
Stat+Conf+RequestDetail
Stat+Exhb+IntDetail&Stat+Exhb+SearchResult&Prop+Guid+Shop
Stat+Exhb+SearchResult&Prop+Guid+Shop
Req+Guid+Shop
Stat+Sel+Shop
Stat+Sel+Shop&Stat+Sel+Shop
Req+Exhb+SearchResult&Stat+Sel+Shop
Stat+Exhb+ShopInfo
Stat+Exhb+IntDetail
Stat+Exhb+IntDetail&Expr+Guid+Shop
Expr+Guid+Shop
Stat+Exhb+SearchResult
Stat+Exhb+IntDetail
None
End

Figure 5: A part of dialogue sequences by the intention tag

4.2 Difference between Human and WOZ

We have recorded in-car information retrieval dialogues with a human navigator, Wizard of OZ, and the ASR system. The ASR system performs a system-initiative dialogue. Therefore, speech styles of subjects for the ASR system are highly restricted from the guidance of the system. In this section, we analyze the difference of subject's behaviors between the human navigator and the WOZ system.

In Figure 6, the number of phrases per speech unit (line) is shown with right vertical pivot for each intention tag. We also investigate the occurrence of linguistic phenomena such as filler for each tag. In Figure 6, we only show the occurrence rate of fillers. The average occurrence of filler is 0.15 per phrase in human dialogue and 0.12 per phrase in WOZ dialogue. From this graph, we can read the dialogue between subjects and the WOZ system is shorter than taht with human on average. This tendency is not affected from LIT. For the "Request(Req)" tags, occurrence rate of fillers is not high and almost average. There are no difference between human and the WOZ, though, the other tags differ with each intention tag. The difference between human navigator and WOZ is also high in the other tags. This means that, for the "Req" tags, subjects usually have an intention to speech and not affected from systems reply. For the other tags, subjects usually reply the systems answer. So the fluency of the system might highly affect the user's speech. Also, from the number of phrases per speech unit, "Req" tagged units are more complex than other tagged units.

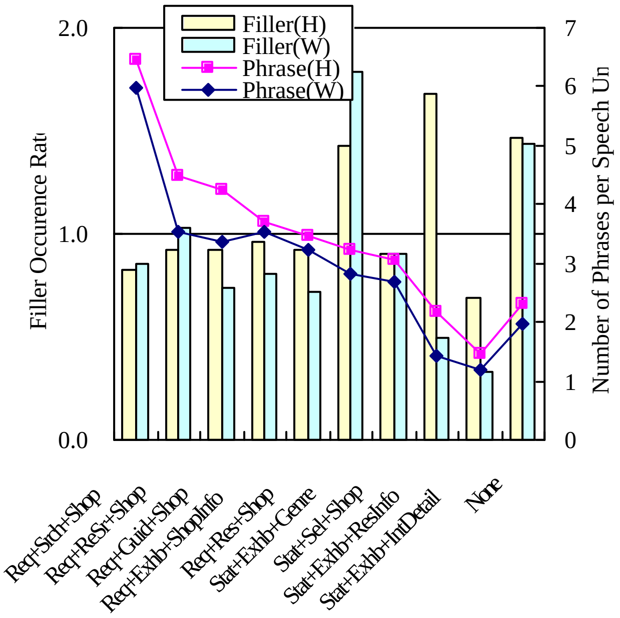

Figure 6: Differences of subject's behaviors
between human and WOZ for each intention tag

5 CONCLUSION

In this paper, we have presented the brief description of a multimedia corpus of in-car speech communication. The corpus consists of synchronously recorded multichannel audio/video signals, driving signals and GPS output. The spoken dialogues of the drivers have been collected under the various styles, i.e., human-human and human-machine, prompted and natural, for the restaurant guidance task domain. An ASR system was utilized for collecting human-machine dialogues.

To date, almost 800 subjects have been enrolled in data collection. All of the spoken dialogues are transcribed with time information. We define the Layered Intention Tag for analysis on the dialogue sequence. Half of the corpus is tagged with LIT. We also attach the structured dependency information to the corpus. By these efforts, the in-car speech dialogue corpus is getting richer and can be recognized as a multi-layered corpus. By utilizing the different layers of the corpus, various analysis of the dialogue can be performed. Currently, we are analyzing the relation between the intention and the occurrence rate of fillers. By using the result of these analyses, we are currently studying the corpus-based dialogue management.

Acknowledgements

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

References

[1] N. Kawaguchi, S. Matsubara, K. Takeda, and F. Itakura: Multimedia Data Collection of In-Car Speech Communication, Proc. of the 7th European Conference on Speech Communication and Technology(EUROSPEECH2001), pp. 2027--2030, Sep. 2001, Aalborg.

[2] D. Roy: "Grounded" Speech Communication, Proc. of the 6 th International Conference on Spoken Language Processing (ICSLP 2000), Vol. IV, pp.69-72, 2000, Beijin.

[3] 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 the 6 th IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'99), pp.393--396 ,1999, Colorado, USA.

[4] H. Murao, N. Kawaguchi, S. Matsubara, and Y. Inagaki: Example-Based Query Generation for Spontaneous Speech, Proc. of the 7th IEEE Workshop on Automatic Speech Recognition and Understanding(ASRU01), Dec.2001, Madonna di Campiglio.

[5] N. Kawaguchi, K. Takeda, S. Matsubara, I. Yokoo, T. Ito, K.i Tatara, T. Shinde and F. Itakura, : CIAIR speech corpus for real world speech recognition, Proc. of the 5th Symposium on Natural Language Processing (SNLP-2002) & Oriental COCOSDA Workshop 2002, pp. 288-295, May. 2002, Hua Hin, Thailand.

[6] N. Kawaguchi, S. Matsubara, K. Takeda, and F. Itakura, Multi-Dimensional Data Acquisition for Integrated Acoustic Information Research, Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Vol. I, pp. 2043-2046, May 2002, Canary Islands.

[7] S. Matsubara, S. Kimura, N. Kawaguchi, Y. Yamaguchi and Y. Inagaki: Example-based Speech Intention Understanding and Its Application to In-Car Spoken Dialogue System, Proc. of the 17th International Conference on Computational Linguistics (COLING-2002), Vol. 1, pp. 633-639, Aug. 2002, Taipei.

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

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

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