CONSTRUCTION AND ANALYSIS OF THE MULTI-LAYERD IN-CAR SPOKEN DIALOGUE CORPUS

Nobuo Kawaguchi, Shigeki Matsubara, Itsuki Kishida, Yuki Irie, Yukiko Yamaguchi, Kazuya Takeda and Fumitada Itakura

Center for Integrated Acoustic Information Research, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8601, JAPAN
http://www.ciair.coe.nagoya-u.ac.jp/

ABSTRACT

In this paper, we report the construction of the multi-layered in-car spoken dialogue corpus and the preliminary result of the 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, multi-channel video data from 3 cameras and the vehicle related data. Multimedia data has been collected for three sessions of spoken dialogue with different types of navigator in about 60-minute drive by each of 800 subjects. We have defined the Layered Intention Tag for the analysis of dialogue structure for each of speech unit. Then we have marked the tag to all of the dialogues for over 35,000 speech units. By using the dialogue sequence viewer we have developed, we can analyze the basic dialogue strategy of the human-navigator. We also report the preliminary analysis of the relation between the intention and linguistic phenomenon.

1. INTRODUCTION

Speech interface which can deal with spontaneous speech is one of the landmarks for the human-machine interface. To attain the landmark, large-scale speech corpora play important roles for both of acoustic modeling and speech 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 acoustic models for the in-car speech data, 2) training language models of spoken dialogue for 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)(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 data has been collected by recording several sessions of spoken dialogue in about a 60-minute drive by each of 800 drivers. All of the spoken dialogues are transcribed with detailed information. We have defined the Layered Intention Tag for analyzing dialogue structure. The data can be used for analyzing and modeling the interactions 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 a car. In Section 3 we introduce the Layered Intention Tag for analysis of dialogue acts. Preliminary studies on analysis of the relation between the intention and linguistic phenomenon are presented in Section 4.

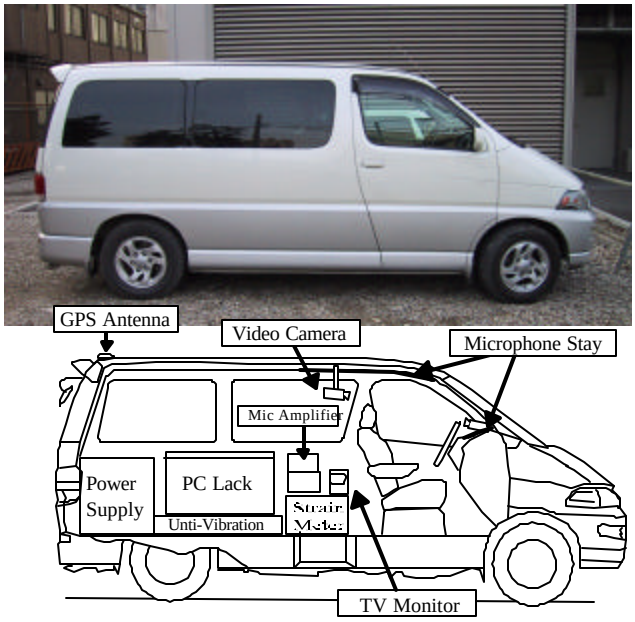

Figure 1: Data Collection Vehicle

Table 1: Collected Speech Data

1999'S COLLECTION	
Spoken dialog 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 dialog with human navigator	5min
Spoken dialog with WOZ system	5min
Spoken dialog with ASR system	5min
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	187.6hour
Sessions	209	589	587	575	1960
Speech len.(sec)	98,100	137,025	98,288	102,933	121.2hour
driver	44,772	54,140	38,286	22,516	44.4hour
operator	53,328	82,885	60,002	80,417	76.8hour
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

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 systems. One is a human navigator, which can talk most fluently and naturally. Another is a WOZ system. Our WOZ system is equipped with a touch panel-PC and speech synthesizer. Figure 2 shows a recording situation of the WOZ system. 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 dialog system with ASR. The system is using 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 Dialog Recording

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)

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
None(-)	Guide(Guid)	SearchResult	Date
	Select(Sel)	RequestDetail	Menu
	Reserve(Res)	SelectionDetail	Count
		YesOrNo	Time

3. LAYERED INTENTION TAG

To develop a spoken dialogue system based on 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 presume the intention of the user's utterances. By the preliminary experience, we learned that user's intention is widely spread even in a simple task. So, if we define the detailed intention tag, we need to define dozens of them. Therefore, we divide the intention tag 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. Some units don't have the tag in this layer. All of Discourse Act tags are "task independent tag". Action layer denotes the action of the speech unit. Action tag is divided into "task independent tag" and "task dependent tag". "Confirm" and "Exhibit" are task independent, but others ("Search", "ReSearch", "Guide", "Select" and "Researve") are task dependent tag. 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 argument layer can be decided directly from the specific keywords in the sentence.

An example of a dialogue between a human navigator and a subject is shown in Figure 3. For each utterance (speech unit), we tagged the LIT. At this time, we have tagged it for over 35,000 speech units.

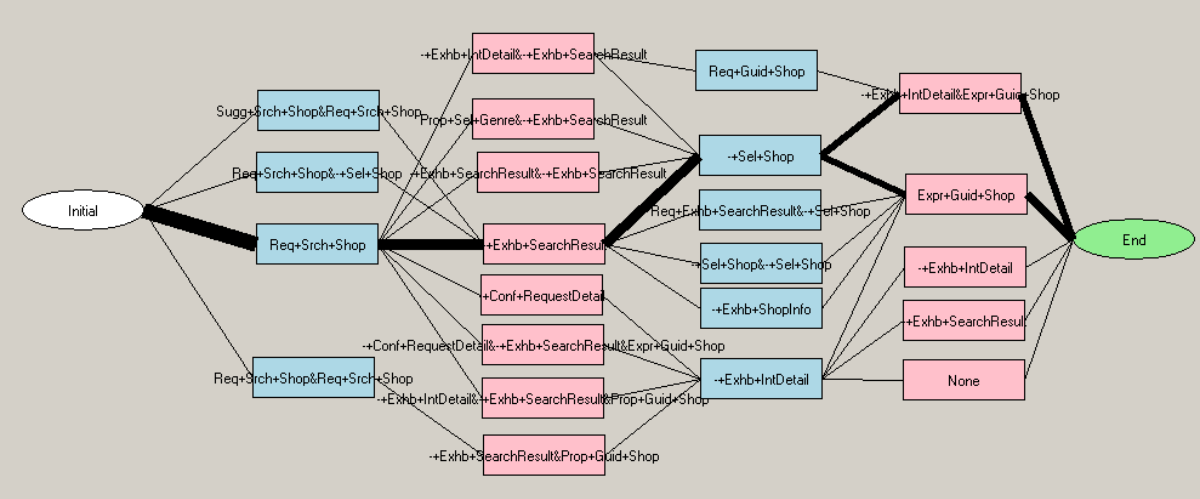

Figure 4: A part of dialogue sequences by the Layered Intention Tag

```
        Utterance        |        LIT
-----------------------------------------------
Subj:Umm, I'm looking for
     a fastfood restaurant.
                              Req +Srch+Shop
Navi:Well, there are McDonald's,
     Mr.Donuts,  and Lotteria near here.
                              - +Exhb+SrchRes
Subj:So, McDonald's please.
                              - +Sel +Shop
Navi:OK. I'll navigate to the
     McDonald's restaurant.
                              Expr+Guid+Shop
```

Figure 3: Example of the Transcription with LIT

4. ANALYSIS OF THE CORPUS

We divide the recording session into the short tasks. Each task is a dialogue about a single theme. The dialogue in Figure 3 is an example of a single task about the restaurant query. We have used provided the tags for the all tasks about the restaurant query. Total number of the tagged task is 3641. For each task, we have 9.7 speech units.

4.1. Dialogue Sequence Viewer

To understand and analyze the dialogue intuitively, we develop a dialogue sequence viewer shown in Figure 4. 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 a occurrence count of the tag's connection. Figure 4 only shows a short dialogue which ends only 4 turns. Average turn count of the restaurant query task is about 10.

By using the dialogue viewer, we found that most of the dialogue sequence pass through the typical tags such as "Req+Srch+Shop", "-+Exhb+SrchRes", "-+Sel+Shop", and "Expr+Guid+Shop". Dialogue in Figure 3 is one of the typical sequences. We also check the dialogue of the length 6, 8 and 10. From this experience, we notice that start section and end section of the dialogue are very similar in different length of dialogues.

4.2. Difference between Human and WOZ

We have recorded in-car information retrieval dialogues with a human navigator, Wizard of OZ, and ASR system. ASR system performs a system initiative dialogue. Therefore, speech styles of subjects for 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 Wizard of OZ system.

In the Figure 5, number of phrases per speech unit (line) is

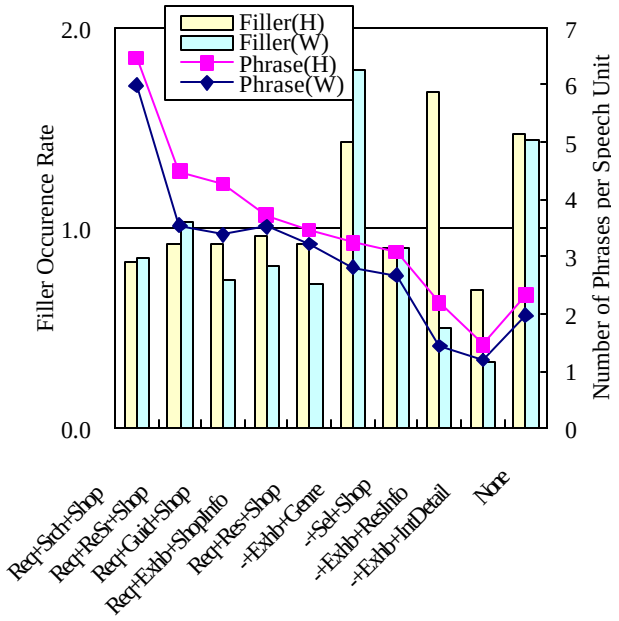

Figure 5: Differences of Subject's Behaviors between Human and WOZ for each LIT

shown with right vertical pivot for each Layered Intention Tag. We also investigate the occurrence of linguistic phenomena such as filler for each LIT. In the Figure 5, we only show the occurrence rate of filler. 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 WOZ is shorter than dialog with human in average. This tendency is not affected from LIT. For the "Request(Req)" tags, occurrence rate of filler is not high and almost average. There are no difference between human and WOZ, though, other tags differ with each LIT. Difference between human navigator and WOZ is also high in 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 fluency of the system might highly affect the user's speech. Also, from the number of phrases per speech unit, "Req" tagged units are most complex sentences than other tagged units.

5. SUMMARY

In this paper, we presented 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 driver were collected in 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 spoken dialogues are transcribed with time information. We define the Layered Intention Tag for analysis of dialogue sequence. Half of the corpus is tagged with LIT. We also attach the structured dependency information to the corpus. By these efforts, in-car speech dialogue corpus is getting richer and can be recognized as a multi-layered corpus. By utilizing different layer of the corpus, various analysis of the dialogue can be performed. Currently, we analyze the relation between LIT and number of phrases and occurrence rate of fillers. By using the result of these analyses, we are currently studying the corpus based dialogue management.

ACKNOWLEGEMENTS

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

11. REFERENCES

[1] Nobuo Kawaguchi, Shigeki Matsubara, Kazuya Takeda, and Fumitada 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] Deb Roy: "Grounded" Speech Communication, Proc. of the International Conference on Spoken Language Processing (ICSLP 2000), pp.IV69--IV72, 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 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'99), pp.393--396 (1999).

[4] Hiroya Murao, Nobuo Kawaguchi, Shigeki Matsubara, and Yasuyoshi 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]Nobuo Kawaguchi, Kazuya Takeda, Shigeki Matsubara, Ikuya Yokoo, Taisuke Ito, Kiyoshi Tatara, Tetsuya Shinde and Fumitada Itakura, ： CIAIR speech corpus for real world speech recognition, Proceedings of 5th Symposium on Natural Language Processing (SNLP-2002) & Oriental COCOSDA Workshop 2002, pp. 288-295, May. 2002, Hua Hin, Thailand.

[6] Nobuo Kawaguchi, Shigeki Matsubara, Kazuya Takeda, and Fumitada Itakura, Multi-Dimensional Data Acquisition for Integrated Acoustic Information Research, Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC-2002), Vol. I, pp. 2043-2046, May 2002, Canary Islands.

[7]Shigeki Matsubara, Shinichi Kimura, Nobuo Kawaguchi, Yukiko Yamaguchi and Yasuyoshi Inagaki ： Example-based Speech Intention Understanding and Its Application to In-Car Spoken Dialogue System, Proceedings 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/