CIAIR IN-CAR SPOKEN DIALOGUE CORPUS AND ITS APPLICATION

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

Center for Integrated Acoustic Information Research (CIAIR), 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 in-car spoken dialogue corpus which has been constructed in CIAIR, Nagoya Univ. We have developed a collecting system specially built in a Data Collection Vehicle (DCV) for synchronous recordings of various kinds of 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 812 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 application of the corpus to the dialogue system

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 our 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 1.8 TB of data has been collected by recording several sessions of spoken dialogue in about a 60-minute drive by each of over 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.

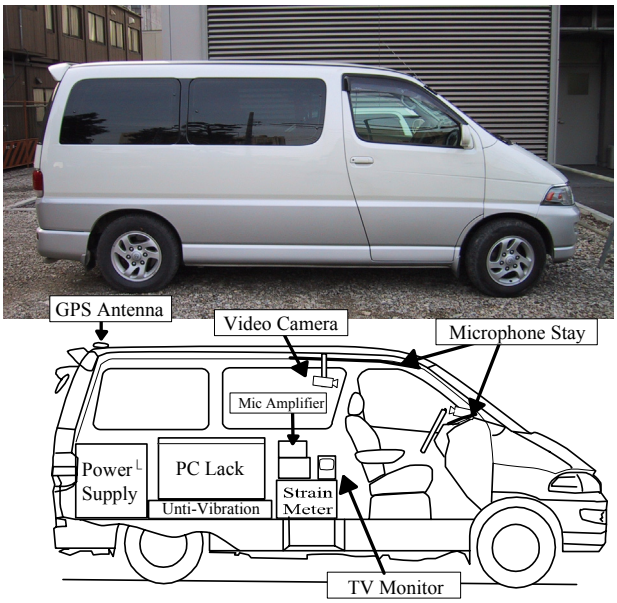

Figure 1: Data Collection Vehicle

In the next section, we describe the multimedia data collection using our Data Collection Vehicle. In Section 3 we introduce the Layered Intention Tag for analysis of dialogue acts. Section 4 briefly describes other layers of the corpus. Preliminary studies on analysis of the relation between the intention and linguistic phenomenon are presented in Section 5.

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: Age distribution of the subjects in the corpus.

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

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: 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

We performed our data collection in CIAIR from 1999 to 2001. We have collected the speech of over 800 subjects while car driving. The collected data is shown in Table 1. At the first year, we had collected 1) pseudo information retrieval dialogue between subject and human navigator, 2) phonetically balanced sentences, 3) isolated words and 4) digit strings for 212 subjects. In 2000-2001's collection, we added 2 more dialogue modes, so that each subject had performed a dialogue with three kinds of systems. First system is a human navigator, which can talk most fluently and naturally. Second one is a WOZ system. The last system is an automatic dialog system with ASR. The system is using Julius [3] for the ASR engine. Age distribution of the corpus is shown in Table 2.

Figure 2: WOZ Dialog Recording

Speech data of read text has been collected from the drivers. Each subject has read 50 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. However, it is dangerous to read a text while driving, subjects are prompted each phonetically sentences from the head-set using special equipped wave-playback software.

A recording 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. The recording system is composed of eight network connected computers, distributed microphones, microphone amplifiers, a video monitor, video cameras, pressure sensors, Differential-GPS, and uninterruptible power supply(UPS). Individual computers are used for speech-input, sound-output, 3 video channels, and vehicle related data. Table 3 shows a specification of the collected data. These multi-dimensional data are recorded synchronously, and can be synchronously analyzed.

2.1. Multi mode dialogue collection

The main concept of the dialogue speech collection is to record three modes of dialogues. The domain of the task is the information retrieval task for all modes. Description of each dialogue mode is as follows.

- Dialogue with human navigator (HUM): He/she gets a workout as a navigator in advance and has the detailed information for the task achievement. However, in order to avoid a dialogue divergence, some restriction is put on the way he/she talks.

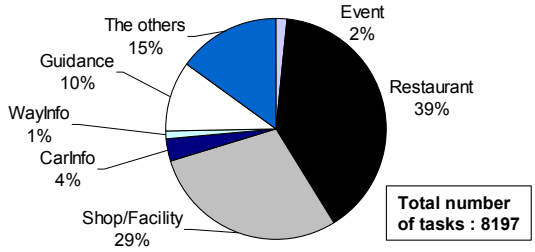

Figure 3: Task distribution of the corpus

- Dialogue with Wizard of OZ system (WOZ): WOZ is a spoken dialogue system which has a touch-panel input by human navigator, and speech synthesizer output. The system has a database of shop and restaurants and the navigator use the system to search and select the suitable answer for subjects utterance (Figure 2).
- Dialogue with Spoken Dialogue System (SYS): The dialogue system called "Logger" performs a slot-filling dialogue for the restaurant retrieval task. The system utilize Julius[3] for LVCSR system.

To ease the dialogue recordings, the navigator have prompted each task by using several level of the task description panel to initiate the spontaneous speech. Examples of the panel are as follows, 'Fast food', 'Hungry', 'Hot summer, thirsty', 'No money', 'You just returned from abroad'. All of recorded dialogues are transcribed into text using criteria for the Corpus of Spontaneous Japanese (CSJ)[13]. Table 4 shows basic statistical information of the dialogue corpus. Finally, we have collected over 187 hour and 1M morphemes dialogues.

We have divided the sessions into tasks. Figure 3 shows the distribution of the task in the corpus. In the corpus collection, we have focused in restaurant search task. So 40% of the tasks are restaurant retrieval task. In the following sections, we only use the restaurant tasks for analysis.

Table 5: 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

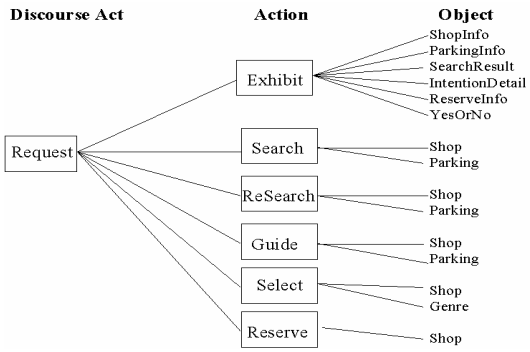

Figure 3: Structure of the Layered Intention Tag(a part of)

3. 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.
                                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

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 5. 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 tags". Action layer denotes the action of the speech unit. Action tag is divided into "task independent tags" and "task dependent tags". "Confirm" and "Exhibit" are task independent, but others ("Search", "ReSearch", "Guide", "Select" and "Reserve") 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. As Figure 3 shows, the lower layered intention tag depends on the upper layered one.

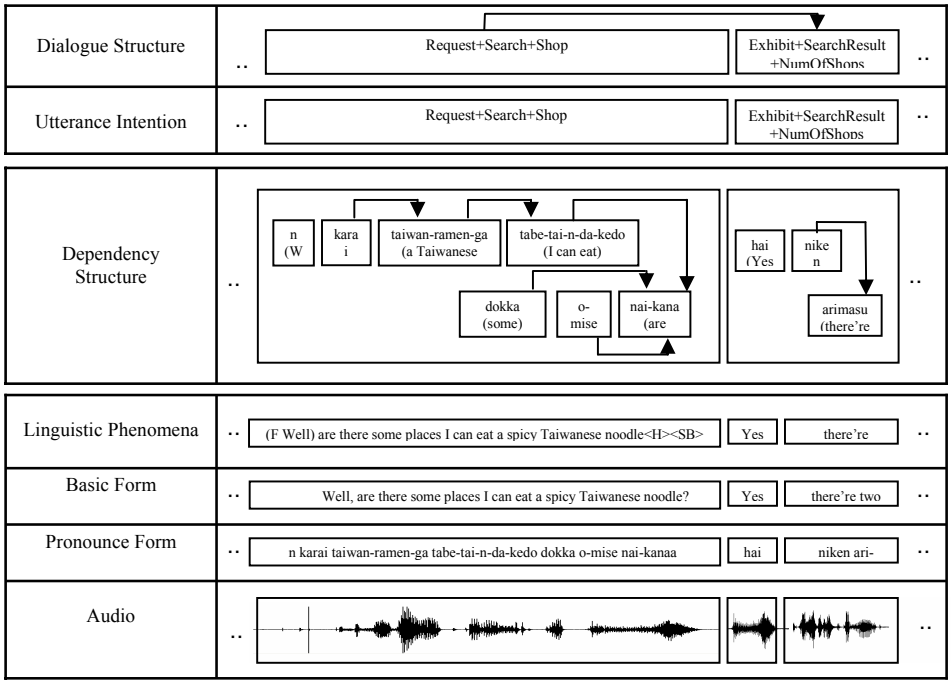

Figure 5: Multi-Layered In-Car speech corpus

```
(TIME 01:48:502-01:54:821)
((1  ( n (Well) filler ))
                    ->   None )
((2  ( karai(spicy) adjective ))
                    ->  (3  ( Taiwanramen-ga (a Taiwanese noodle) noun-particle )))
((3  ( Taiwanramen-ga (a Taiwanese noodle) noun-particle ))
                    ->  (4  ( tabe-tai-n-da-kedo (I can eat) verb-auxiary-noun-auxiary-particle )))
((4  ( tabe-tai-n-da-kedo (I can eat) verb-auxiary-noun-auxiary-particle ))
                    ->  (7  ( nai-ka-na (are there) adjective-auxiary-auxiary )))
((5  ( dokka (some) noun ))
                    ->  (7  ( nai-ka-na (are there) adjective-auxiary-auxiary )))
((6  ( o-mise (places) prefix-noun ))
                    ->  (7  ( nai-ka-na (are there) adjective-auxiary-auxiary )))
((7  ( nai-ka-na (are there) adjective-auxiary-auxiary ))
                    ->  None )
```

Figure 6: An example of the corpus with dependency tags.

4. MULTI-LAYERED STRUCTURE

Generally, a spoken dialogue system can be developed by the combination of the different level of components, such as speech processing, language processing, dialogue processing, and so on. In order to use the collected dialogue data for upgrading a system, not only a simple recording and transcription of speech but advanced information is needed. Then, we have advanced the dialogue corpus by giving various linguistic analyses on syntax and semantics to the text data of the corpus. Thereby, the multi-layered spoken dialogue corpus as shown in Figure 5 could be constructed.

4.1. Corpus with Dependency Tags

We gave the dependency analysis to the driver's utterances. Dependency in Japanese is a dependency relation between the head of bunsetsu and the other bunsetsu. In addition, the bunsetsu, corresponding to the basic phrase in English roughly, is about the relation between an utterance intention and utterance length, and the relation between utterance intentions and linguistic phenomena. Especially, paying attention to the minimum unit into which a sentence can be divided naturally in terms of meanings and pronunciations. The dependencies might be over two utterance units which are segmented by a pause. And such a dependency as a bunsetsu depends on a forward bunsetsu is also accepted. So we adopt the data specification accommodating to spontaneous utterances. The example of a corpus with dependency tags is shown in Figure 6. The corpus includes not only the dependency between bunsetsus but morphological information, utterance unit information, dialogue turn information, and so on. Thus it has various levels of the linguistic information. This corpus is used for acquisition of the dependency probability for stochastic dependency parsing[14].

5. ANALYSIS OF THE CORPUS

We gave the characteristic analysis to the advanced spoken dialogue corpus. This section describes the result of the analysis about the relation between an utterance intention and utterance length, and the relation between utterance intentions and linguistic phenomena. Especially, paying attention to the driver's utterances in human-human conversations and human-WOZ conversations, we compare those utterances.

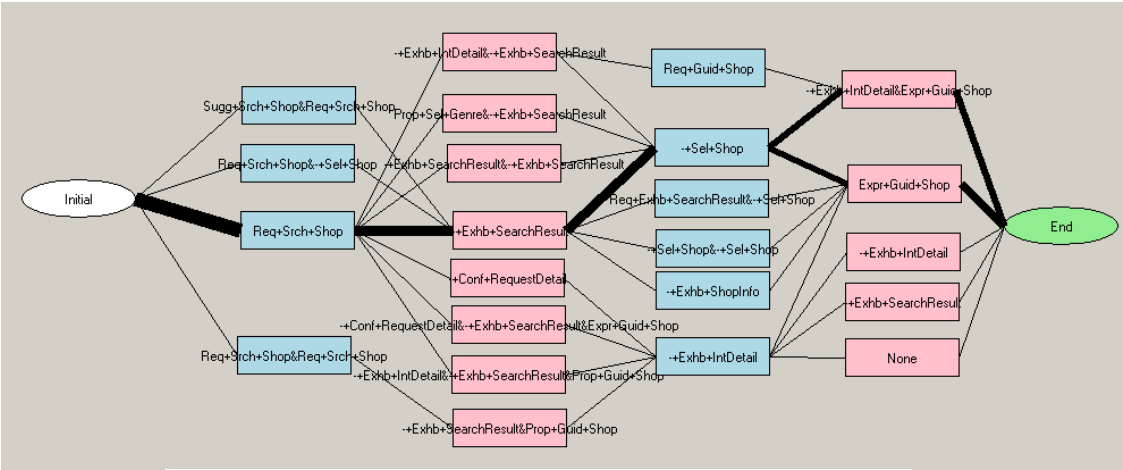

Figure 7: A part of dialogue sequences by the Layered Intention Tag

5.1. Dialogue Sequence Viewer

To understand and analyze the dialogue intuitively, we develop a dialogue sequence viewer shown in Figure 7. 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 7 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","Stat+Exhb+SrchRes","Stat+Sel+Shop", and "Expr+Guid+Shop". 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 start section and end section of the dialogue are very similar in different length of dialogues.

5.2. 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 8, number of phrases per speech unit (line) is 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 8, we only show the occurrence rate of filler.

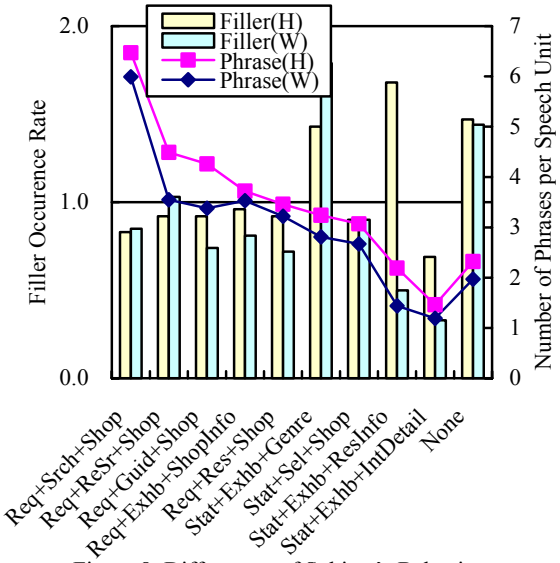

Figure 8: Differences of Subject's Behaviors between Human and WOZ for each LIT

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.

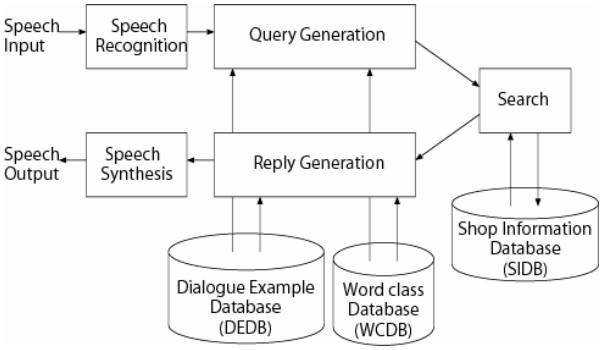

Figure 9: Configuration of example-based dialogue system.

6. APPLICATION OF THE CORPUS

As one of the applications of the corpus, we have built a example based spoken dialogue system which can utilize a WOZ system log(Figure 9)[4]. The system uses the WOZ system log to determine the system behavior to the subject's utterance using closest examples.

7. SUMMARY

This paper presents a description of the multimedia corpus of in-car speech communication developed in CIAIR. The corpus consists of synchronously recorded multi-channel 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.

Finally, more than 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.

ACKNOWLEDGEMENT

This work was supported in part by a Grant-in-Aid for COE Research (No. 11CE2005) of the Ministry of Education, Science, Sports and Culture, Japan. The authors would like to thank all members of CIAIR for their great contribution and large efforts to the construction of the in-car spoken dialogue corpus.

REFERENCES

[1]Nobuo Kawaguchi, Shigeki Matsubara, Kazuya Takeda, and Fumitada Itakura: Multimedia Data Collection of In-Car Speech Communication, EUROSPEECH2001, pp. 2027--2030,(2001).

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

[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, Yukiko Yamaguchi and Yasuyoshi Inagaki: Example-based Spoken Dialogue System using WOZ System Log, Proc. of The 4th SIGdial Workshop on Discourse and Dialogue (SIGDIAL2003),pp.140—148(2003).

[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, Oriental COCOSDA Workshop 2002, pp. 288-295(2002).

[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(2002).

[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, Proc. of the 17th International Conference on Computational Linguistics (COLING-2002), Vol. 1, pp. 633-639, (2002).

[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, EUROSPEECH2001, pp. 2023--2026,(2001).

[9]P. A. Heeman, D. Cole, and A. Cronk : The U.S. SpeechDat-Car Data Collection, EUROSPEECH2001, pp. 2031--2034, (2001).

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

[11]Yuki Irie, Nobuo Kawaguchi, Shigeki Matsubara, Itsuki Kishida, Yukiko Yamaguchi, Kazuya Takeda, Fumitada Itakura, and Yasuyoshi Inagaki: An Advanced Japanese Speech Corpus for In-Car Spoken Dialogue Research, in Proc. of Oriental COCOSDA-2003, pp.209− 216(2003).

[12]Itsuki Kishida, Yuki Irie, Yukiko Yamaguchi, Shigeki Matsubara, Nobuo Kawaguchi and Yasuyoshi Inagaki: Construction of an Advanced In-Car Spoken Dialogue Corpus and its Characteristic Analysis, EUROSPEECH2003, pp.1581—1584(2003).

[13]K. Maekawa, H. Koiso, S. Furui, H. Isahara, "Spontaneous Speech Corpus of Japanese", LREC-2000, No.262(2000).

[14]S. Matsubara, T. Murase, N. Kawaguchi and Y. Inagaki,"Stochastic Dependency Parsing of Spontaneous Japanese Spoken Language", Proc. of the 17th International Conference on Computational Linguistics (COLING-2002), Vol.1, pp.640-645(2002).