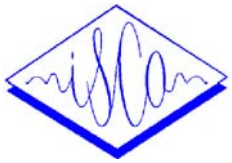

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A SPOKEN DIALOG SYSTEM FOR A MOBILE OFFICE ROBOT

Hideki Asoh, Toshihiro Matsui, John Fry†, Futoshi Asano, Satoru Hayamizu

Real-World Intelligence Center
Electrotechnical Laboratory
1-1-4, Umezono, Tsukuba
Ibaraki 305-8568, JAPAN
jijo2@etl.go.jp

†Linguistics Dept. and CSLI
Stanford University
220 Panama Street
Stanford, CA94305-2150, U.S.A.
fry@csli.stanford.edu

http://www.etl.go.jp/~7440/

ABSTRACT

A spoken dialog interface of a mobile office robot is described. To realize robust speech recognition in noisy office environments, a microphone array system and a technique of switching multiple speech recognition processes with different dictionaries are introduced. To realize flexible and natural dialog, task dependent semantic frames and keeping track of attentional state of dialog are used. The system is implemented on a real mobile robot and evaluated with sample dialogs.

Keywords: Spoken Dialog Interface, Office Robot

1. INTRODUCTION

We have been developing a learning mobile office robot called *Jijo-2* (Figure 1)[1][2][8]. The robot is expected to provide services such as guidance for visitors, delivery of messages, office members' schedule management, meeting arrangement, and so on. That is, its tasks are management and presentation of office-related information. Because real offices are dynamically changing environments, in order to accomplish the tasks, the robot should autonomously gather information through multi modal sensing and remote database access, and continue to learn about the office. Making dialog with office members is also an important way to get information. Hence, for the robot, spoken dialog interface is crucial not only for naive users to gain easy access to the robot's services, but also for the robot itself as a powerful tool for gathering information.

Although the idea of making dialog with robot is rather old[10], and the recent development of speech recognition and natural language processing technologies are remarkable, only a few spoken dialog systems have been actually implemented on real robots[5][11], and the dialog capability of implemented systems is not satisfiable. Most of them are used just for giving simple commands to the robot.

Three major problems, which we encounter in developing spoken dialog systems for robots, are

1) realization of robust speech recognition for noisy environments,
2) realization of real-time responses with limited computational resources, and
3) realization of flexible dialog control covering a wide variety of behaviors of the robots.

In this paper, in order to show our current solution to the above problems, we first briefly describe the dialog system for our mobile office robot *Jijo-2*. More details of the system can be seen in our previous papers[4][9]. Then we discuss features and problems of our system.

Figure 1: *Jijo-2* is talking with a user.

2. SYSTEM OVERVIEW

A spoken dialog system is normally composed of three parts. The first part is a speech understanding module, which decodes input speech signal to semantic contents. The second part manages dialog process and executes problem solving using the contents and system's knowledge. The third part encodes system's intention into speech. In our system, the first part is composed of a microphone array system for beam forming and noise reduction, a speaker independent continuous Japanese speech recognizer, and a simple syntactic and semantic analyzer. The second part is realized with a rule-based state transition network with additional mechanisms for maintaining the state of dialogs. The third part uses prepared reply templates with slots filled by the context of dialogs. The whole structure of the system is shown in Figure 2.

2.1 Microphone Array and Vector Quantizer

The microphone array system uses eight omnidirectional microphones. The microphones are arranged around a half circle. The diameter of the circle is about 25 cm, and the distance between neighboring microphones is about 6 cm. The signal from each microphone

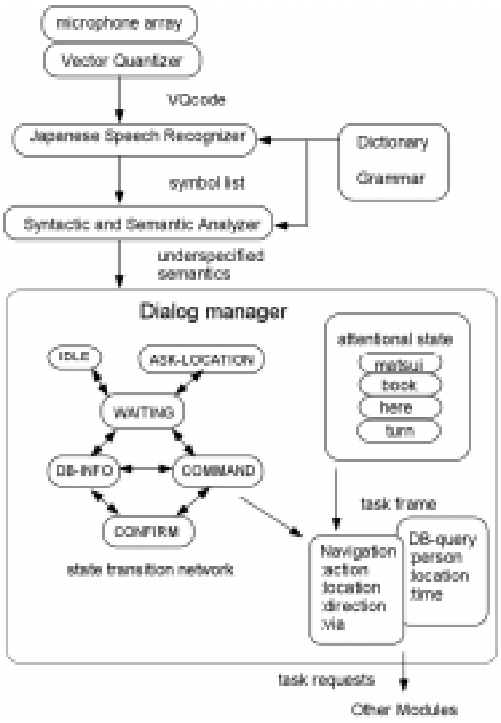

Figure 2: System Overview

is digitized simultaneously and fed into a DSP (TI TMS320C44). The DSP executes identification of sound source direction and reduction of ambient noises using delay-and-sum beam forming algorithm. Direction identification is executed approximately every second while a speech signal is detected. Noise reduction of about 10 dB was obtained in the range of 1500 to 3000 Hz by using the system.

The output of microphone array system is sent to a vector quantizer. Another DSP (C44) is used for the frequency analysis and the vector quantization. The VQ code for each phonetic element is emitted every 10msec. The codes are buffered and sent to the CPU of the robot through slow serial link operating at 10Hz.

2.2 Decoder

The sequence of phonetic VQ codes are decoded by a speaker-independent continuous Japanese speech recognizer, *NINJA*, developed at our laboratory[6]. *NINJA* uses HMM-based phoneme models, a word dictionary, and a context free grammar to decode symbols into a sentence. We prepare a specific word dictionary and grammar tuned for the office robot tasks. The dictionary contains about 200 words including names of office members and the grammar has about 90 generation rules.

In order to make speech recognition robust, we also employ two other simpler grammars, which are used in specific states of dialogs. The first is *reply-grammar*, which accepts just reply sentences denoting "yes" or "no". The other is *name-grammar*, which accepts only sentences denoting names of locations or persons. Three decoding processes with different dictionaries run concurrently, and an output from a decoder is taken up by the following the other modules depending on the state of dialog.

```
sentence: greeting
sentence: imperative
sentence: declarative
sentence: interrogative
imperative: action
imperative: action please
action: motion
action: direction to motion
direction: RIGHT
direction: LEFT
```

Figure 3: Excerption of the Grammar

Since audio beam forming, vector quantization, and decoding are performed in a pipelined manner on three processors, the whole process finishes almost in real-time. The largest latency is 0.4 second for the pause detection to identify the end of an utterance.

2.3 Syntactic and Semantic Analyzer

To understand the intention of a speaker and to respond correctly, the semantic content of a utterance should be extracted from the result of speech recognition. For this purpose, a parser and a semantic dictionary are normally used. Instead of using a standard, general purpose parser and semantic dictionary, we chose to use the simple task-dependent dictionary and grammar from the speech recognizer to parse the recognition result. We designed the dictionary and the grammar in order to execute semantic analysis at the same time. The main reasons are 1) the general parser often fails to parse conversational Japanese utterances with recognition errors, 2) computational cost of the parser is too expensive to realize delay-less responses, and 3) detailed syntactic information is not necessary in the following processes.

In this word dictionary, we embed task-dependent semantic equivalences. For example, "Ohayo (Good morning)," "Konnichiwa (Good afternoon)," and "Konbanwa (Good evening)" all output the same symbol "hello". In the grammar, instead standard tokens representing grammatical categories, we embed tokens representing task-dependent semantic categories. In Figure 3, a part of the grammar is depicted, where "action" and "direction" are such tokens.

Using the dictionary and grammar, "Ohayo (Good morning)" is recognized as (hello) and then parsed and translated into (greeting hello). "Migi ni magatte (Turn to the right)" is recognized as (right to turn) and then translated into (imperative :action turn :direction right). "Matsui-san wa doko? (Where is Dr.Matsui?)" is recognized as (Matsui wa where) and translated into (interrogative :person matsui :question location).

2.4. Dialog Manager

Outputs of the parser are sent to the dialog manager. The dialog manager maintains the state of the dialog and outputs appropriate responses from the robot. In order to maintain the state of the current dialog, the manager uses the combination of three schemes of representing the state. The first is a state transition network where the state is represented as a current state in the finite automaton network. Depending on the state, the robot's re-

sponses to input utterances change. The current system has six states, IDLE, WAITING, DB-INFO, COMMAND, ASK-LOCATION, and CONFIRMATION. The rules for response generation and state transition are written in Prolog like logical statements.

The second is the "task frame", which is a kind of a frame structure with several slots to represent and store information necessary to execute a specific task. Currently we define five kinds of task frames: DB-query, DB-update, identify-person, navigation, and call-person. A task frame is created when the system recognizes that the speaker is submitting a new task and is used to accumulate information distributed over several short utterances. When a new descriptor is created, the information contained in the utterance which invoked the creation is used to fill the corresponding slots. For some slots, predefined default values are used to fill them. The information held in the attentional state is also used to fill slots.

The third device is "attentional state" which is a list of salient entities referred to in the preceding utterances. In natural Japanese conversations, salient information which can be easily gleaned from context is not repeated and often omitted from the utterance. Hence, in order to understand under-specified sentences, keeping track of salient information in a conversation is necessary. For the purpose we introduced a short list of salient discourse entities (the people, objects, and events being talked about). The mechanism to maintain the contents of the list is based on the "centering" theory of linguistics[12] and is discussed in our previous paper [4].

3. EXAMPLE DIALOG

Figure 4 illustrates an example of dialog between users and the robot including several different kinds of behaviors. During the dialog, the robot's behavior includes 1) turning to the sound source (user1), 2) detecting and recognizing face of the user, 3) referencing the database of the office members' current location, 4) calling a member by sending e-mail, 5) guiding the user to a member's office, and 6) registering new location information into the location database.

When the system starts up, the dialog manager is put in the IDLE state. In the state the robot accepts only greeting sentences. When a greeting is heard, the dialog manager (1) submits a request to turn in the direction of sound source to the movement control module, (2) requests detection and recognition of the speaker's face from the vision control module, (3) requests the output of "Konnichiwa (Good afternoon)" from the speech synthesizer module, and finally (4) shifts the current state to WAITING.

The next utterance of the user is a question asking about the location of an office member. Here, a DB-query task frame is created. All necessary slots are filled with the information in the utterance. An inquiry to the location database implemented with PostgresSQL on a remote workstation is made through the database manager module of the robot, and the returned result is reported to the user.

U1(Matsui): Kon-nichiwa. (Hello.)
The robot turns to the talker, detects the face, and recognize.
R: Konnichiwa, Matsui-san. (Hello, Dr.Matsui.)
U1 Asoh-san wa doko? (Where is Mr.Asoh?)
R: Honkan ni imasu. ([He is] in the main building.)
U1: Yonde kudasai. (Please call [him].)
R: Asoh-san ni me-ru wo okurimashouka?
(Shall I send E-mail to Mr.Asoh?)
U1: Hai. (Yes.)
U1: Motomura-san wa? ([Where is] Mr. Motomura?)
R: Office ni imasu. ([He is] in [his] office.)
U1: Tsureteitte kudasai. (Please take me [to his place].)
R: Motomura-san no heya ni yukimasu.
([I'll] go to Mr. Motomura's office.)
The robot plans a path to the office and executes the plan. During the movement, it loses its position.
R: Koko wa doko desuka? (Where am I?)
U1: Hara-san no heya no mae desu.
([You are] in front of Dr.Hara's office.)
R: Arigatou. (Thanks.)
The robot confirm the position, re-plans the path, and continues the movement.
R: Motomura-san no tokoro ni tsukimashita.
([We are] at Mr.Motomura's office.)
U2 (Motomura): Shokuji ni yukimasu.
([I'm] going out for lunch.)
R: Motomura-san wa shokuji de yoroshii desuka?
(Is Mr.Motomura going to lunch?)
U2: Hai. (Yes.)
R: Schedule wo touroku shimasu.
([I'll] register [the data] to the database.)
U1: Matsui mo. (Matsui, too.)
R: Matsui-san wa shokuji de yoroshii desuka?
(Is Dr.Matsui going out for lunch?)
U1: Hai (Yes.)
R: Schedule wo touroku shimasu.
([I'll] register [the data] to the database.)

Figure 4: Example dialog
"U" means user and "R" means robot.

The example also illustrates how the omitted salient information is reconstructed. The words in square brackets in the English translation of the Japanese utterance are omitted in Japanese. For example, in the next utterance, just a name of person is told by user1. Because the utterance is also question, the system creates a DB-query frame, and fill a slot using the information "doko (where)" in the attetional state. Such estimation of salient information can be seen everywhere in the dialog.

In the middle of the dialog, during the movement to Mr. Motomura's office, the robot happens to become uncertain about its position. This can occur because of noises in sensor signals. The driver module recognizes the problem and sends the request of asking location to the dialog manager. The robot actively starts a new conversation to confirm its location.

4. EVALUATION AND DISCUSSION

We have not yet done quantitative performance evaluation of the whole dialog system. We demonstrated the system to visitors over 50 times so far. In each demon-

stration, a short dialog like the example was executed. Here we describe our impression from the demonstration experiences and discuss about the features and problems of our system.

Very roughly speaking, about 80% of users' utterances are recognized correctly. The state transition network effectively eliminates wrong responses to spurious utterances generated by noises and prevents the system from catastrophic faults. In almost all demonstrations, even some utterances are miss-recognized, tasks are accomplished by repeating or correcting the utterances.

Introducing the task frame reduces the complexity of the state transition network. Had it not been for the frame structure, we should have prepared many states representing the state of the slot filling and many state transition rules. Using the structure, the dialog process can be guided simply by a general rule for slot filling. The task frame is also useful to manage multiple dialogs concurrently. There are cases when a new task submission is interruptively started during the slot filling of other task frames. In such cases, a new task frame is created and the slot filling of the newer frame is given a higher priority.

Differ from usual dialog systems where only human users starts conversations, a unique feature of our system is that the robot itself has strong motivations to start a new conversation. In order to realize this "bi-directional" conversation, we introduced an event-driven multi-agent software architecture where several robot control modules run concurrently[3][7]. Each module monitors interesting events in the environment and can invoke a new conversation depending on the events if it is necessary.

The biggest problem of our system is that the design of the state transition network and task frames are rather ad hoc and strongly task dependent. The scheme of representing semantic information of the utterance is also informal. These problems make it difficult to extend the system to cover wider tasks. When a new task is introduced, large part of the structures must be re-designed. To avoid the problem and make incremental extension easier, designing more systematic semantic representation is necessary. For the design, we should understand about the robot's tasks more deeply and investigate the ontology of the task domains.

5. CONCLUSION

A dialog system for a mobile office service robot is described. In order to make the speech recognition robust, a microphone array and a technique of switching multiple speech recognition processes with different dictionaries are introduced. In order to get real-time response of the system, two DSPs are used for beam forming of the microphone array and for vector quantization of the speech signal. A simple semantic analyzing scheme is also effective. For realizing flexible and natural dialog covering several kinds of tasks of the robot, a frame structure for accumulating information which is necessary to execute a task and attentional state which maintains salient information in a conversation are implemented.

The effectiveness of the techniques is confirmed through demonstrations with example dialogs. Future work includes quantitative evaluation of the system, systematic refinement of the dialog modules based on studies of dialog corpora, and extending the system to cover new tasks.

6. REFERENCES

[1] Asoh, H., Y. Motomura, T. Matsui, S. Hayamizu, and I. Hara (1996), Combining Probabilistic Map and Dialogue for Robust Life-long Office Navigation. *Proceedings of the 1996 IEEE/RSJ International Conference on Intelligent Rbots and Systems*, pp.807-812.

[2] Asoh, H., S. Hayamizu, I. Hara, Y. Motomura, S. Akaho, and T. Matsui (1997), Socially Embedded Learning of the Office-Conversant Robot *Jijo-2*, *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pp.888-885.

[3] Asoh, H., I. Hara, and T. Matsui (1998), Dynamic Structured Multi-agent Architecture for Controlling Office-Conversant Mobile Robot, *Proceedings of 1998 IEEE International Conference on Robotics and Automation*, pp.1552-1557.

[4] Fry, J., H. Asoh, and T. Matsui (1998), Natural Dialogue with the *Jijo-2* Office Robot. *Proceedings of the 1998 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.1278-1283.

[5] Hasimoto, S. et al. (1997), Humanoid Robot -Development of an Information Assistant Robot *Hadaly*, Proceedings of the Sixth IEEE International Workshop on Robot and Human Communication.

[6] Itou, K., S. Hayamizu, K. Tanaka, and H. Tanaka (1993), System Design, Data Collection and Evaluation of a Speech Dialogue System. *IEICE Transactions on Information and Systems*, vol. E76-D, pp.121-127.

[7] Matsui, T., H. Asoh, and I. Hara (1997), An Event-driven Architecture for Controlling Behaviors of the Office Conversant Mobile Robot *Jijo-2*. *Proceedings of the1997 IEEE International Conference on Robotics and Automation*, pp.3367-3371.

[8] Matsui, T., and *Jijo-2* Group (1999), *Jijo-2* Project Web home pages, http://www.etl.go.jp/~7440/.

[9] Matsui, T., H. Asoh, J. Fry, Y. Motomura, F. Asano, T. Kurita, I. Hara, and N. Otsu (1999), Integrated Natural Spoken Dialog System of *Jijo-2* Mobile Robot for Office Services, *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, to appear.

[10] Nilsson, N. (1969), A Mobile Automaton: an Application of Artificial Intelligence Techniques. *Proceedings of the First International Joint Conference on Artificial Intelligence*, pp.509-520.

[11] Takeda, H., N. Kobayashi, Y. Matsubara, and T. Nishida (1997), Towards Ubiquitous Human-robot Interaction. In *Working Notes for IJCAI-97 Workshop on Intelligent Multimodal Systems*, pp. 1-8.

[12] Walker, M., M. Iida, and S. Cote (1994), Japanese Discourse and the Process of Centering. *Computational Linguistics*, vol. 20, pp.193-233.