Integrated Natural Spoken Dialogue System of Jijo-2 Mobile Robot for Office Services

Toshihiro Matsui Hideki Asoh John Fry† Youichi Motomura

Futoshi Asano Takio Kurita Isao Hara Nobuyuki Otsu

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

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

Abstract

Our Jijo-2 robot, whose purpose is to provide office services, such as answering queries about people's location, route guidance, and delivery tasks, is expected to conduct natural spoken conversation with the office dwellers. This paper describes dialogue technologies implemented on our Jijo-2 office robot, i.e. noise-free voice acquisition system by a microphone array, inference of under-specified referents and zero pronouns using the attentional states, and context-sensitive construction of semantic frames from fragmented utterances. The behavior of the dialogue system integrated with the sound source detection, navigation, and face recognition vision is demonstrated in real dialogue examples in a real office.

Motivation

To extend applications of robots from routine tasks in traditional factories to flexible services in the offices and homes, robots are expected to have better man-machine interfaces for two reasons: (1) such settings are more dynamic and therefore require greater flexibility and adjustment to the environment on the part of the robot, and (2) the users are usually non-experts without special training in controlling or programming the robot.

We have been building a learning mobile office robot *Jijo-2* which navigates in an office, interacts with its environment and people, and provides services such as answering inquiries about people's location, route guidance, and delivery tasks. In such systems, natural man-machine interface, especially a spoken dialogue capability, plays an important role. This capability benefits not only the office users but also *Jijo-2* itself, since actions to ask nearby people for help can be taken. For example, we previously implemented a dialogue-based form of map learning and robust navigation where the robot asks a human trainer about its location when it loses its way (Asoh et al. 1996).

Our next challenge focused on realization of more natural dialogue to provide wider range of office services. For such office dialogues, we planed to add query/update to the location and schedule database of laboratory members and task commands such as Email and FAX as well as navigation dialogues for map learning. From the viewpoint of dialogue management, this extension incurs at least two problems:

Figure 1: Jijo-2 robot is talking with a human user while it is navigating in an office corridor.

A: degradation of speech recognition performance due to larger vocaburary

B: composition of semantics from many fragmented utterances, some of which are often omitted.

For **A**, we have employed the following techniques:

1. Multiple microphone array to extract clean voice signal even in noisy offices
2. Multiple voice recognizer processes employing different dictionaries and grammars to allow selection of the most plausible interpretation according to contexts.

The problem **B** is important to make Japanese conversation natural. Even a simple schedule registration can become fairly long, for example, "Matsui, a member of Jijo-2 group, will join a meeting in the main building from 3 to 5 tomorrow". We cannot expect every user to speak such a long sentence in one breath. A natural utterance is usually very short and total meaning is conveyed as a series of utterances. A shorter utterance is also preferred because of its better chances of being correctly recognized. In order to compose

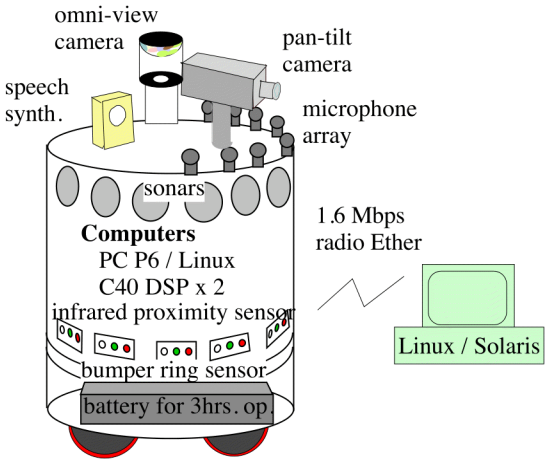

Figure 2: Hardware configuration of Jijo-2.

a complete meaning from fragmented utterances, we introduced the *dialogue context*, which is a kind of frame.

Once we allow split of a long sentence into several short utterances, another problem comes in: natural Japanese has an obvious tendency to omit subject, object, or any pronouns from an utterance whenever it is clear from context. To cope with this problem, we incorporated the *attentional state manager* to understand under-specified statements.

The major topic of this paper is the implementation of the Japanese dialogue subsystem in the *Jijo-2* office mobile robot designed to utilize robot's navigation and database connectivities. The paper is organized to reflect the order in which the dialogue components are used to recognize an underspecified human utterance in a noisy environment, assemble a semantic representation, and finally perform the required behaviors.

Overview of the Jijo-2 Architecture

Jijo-2 is based on the Nomad 200 mobile robot platform manufactured by *Nomadic Inc.* It is equipped with various sensors such as ultrasonic range sensors, infrared proximity sensors, tactile sensors, and an odometric sensor (Figure 2). The on-board computer is a PC running Linux and connected to a LAN through radio Ethernet. We added a microphone array, two CCD color cameras, digital signal processors (DSP) for processing sound signal, and a Japanese speech synthesizer "Shaberimbo".

On the robot and the remote host we have implemented several software modules for navigation and dialogue (Asoh et al. 1996; Matsui, Asoh, and Hara 1997). The overall structure of modules including *reactor modules* and *integrator modules* is depicted in Figure 3. The integrator modules are implemented in *EusLisp*, an object oriented Lisp for robot control (Matsui and Hara 1995), and reactor modules are in C for the sake of realtime control. The modules are managed in an event-driven architecture and realize both reactive and deliberative behaviors (Matsui, Asoh, and Hara 1997). Communication between modules takes place over TCP/IP

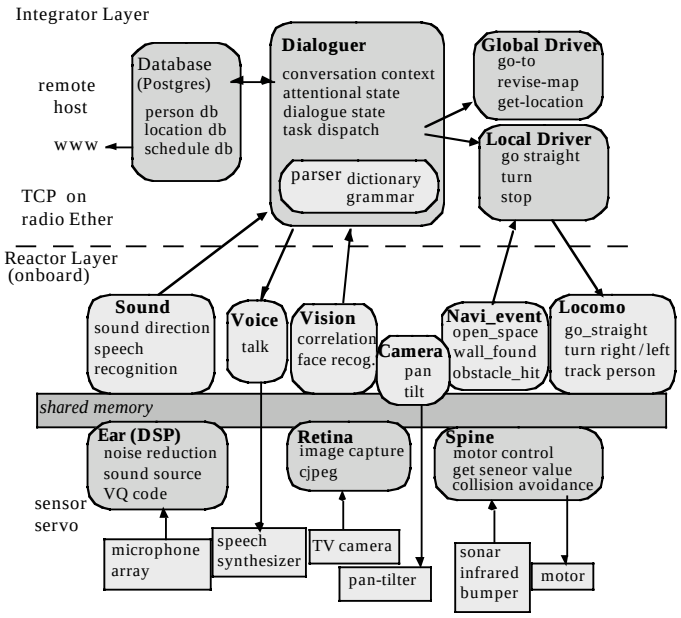

Figure 3: Organization of software modules of Jijo-2.

connections. Major advantages of this event-driven multi-agent architecture are implementation of concurrent behaviors (Figure 1) and plug-and-play of software modules.

Dialog System

Sound Input with Beam-Forming

In order for the *Jiio-2* to conduct smooth dialogue in real environments, dynamic noise suppression to keep a good speech recognition performance is needed. We applied a microphone array composed of eight omni-directional microphones around the top tray of the robot. Sound from a speaker arrives at each microphone with different delays. Sound signal at each microphone is digitized and fed to the first DSP (TI-C44). Based upon the delay-and-sum beam forming method (Johnson and Dungeon1993), the direction to the sound source is computed, which is then used to form a beam to pick up the speech and reduce the ambient noises. We observe noise reduces approximately by 10dB at 3000Hz, which is crucial to the recognition ofconsonants. Figure 4 shows our multi-microphone system can keep relatively good performance even in noisy environments compared with a single microphone system.

Japanese Speech Recognition

The noise-clean digital sound data is sent to the second DSP through the DMA-driven communication port. The second DSP does the frequency analysis and emits the vector quantization code (VQ code) for each phonetic element every 10 ms. The VQ codes are sent to the on-board PC through a serial link.

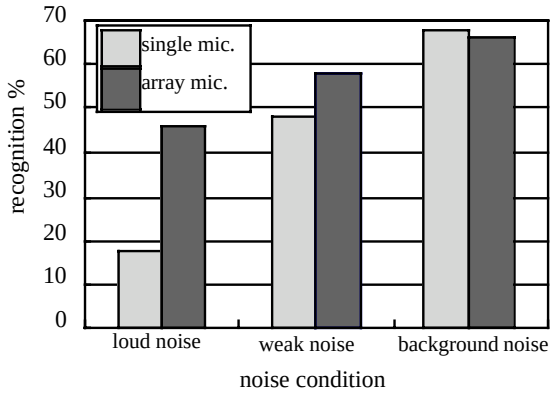

Figure 4: Speech recognition performance of multi-microphone array in various noise conditions.

The continuous speaker-independent Japanese speech recognizer, called *NINJA*, is a HMM-based system developed at our laboratory (Itou et al. 1993). Using a phonetic dictionary created by the HMM learning beforehand, the *NINJA* searches for a series of word symbols that satisfy a grammar. Thus, the speech recognition module produces a list of symbols such as, *(hello)*, *(right to turn)*, *(straight to go)*, *(here is Matsui s office)*, etc., together with a recognition confidence value and a direction angle to the sound source (Note that objects precede verbs in Japanese).

Since the audio beam forming, VQ-code generation, and speech recognition are performed by the pipelined multi-processors, the total recognition finishes almost in real-time. The greater part of a latency is brought by the pause for 0.4 second to identify the end of an utterance.

When the robot starts up, three speech recognition processes begin execution. Each recognizer handles one grammar, i.e. *reply* grammar, *location* grammar, and *full* grammar. The dialogue manager in the integrator layer chooses one grammar at a time. When a yes-or-no reply is expected, the *reply* grammar is activated. When location/person names are expected, the *location* grammar is invoked. Otherwise, the *full* grammar is used. This scheme has proven useful in raising the success rate of recognition.

Parsing

The grammar employed by the speech recognition roughly has the following descriptions:

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

Phonetic representations of the terminal symbols like `RIGHT and LEFT` are given separately in lexicon definition descriptions. These representations can be interpreted as not only deductive rules that expand to instances of acceptable sentences, but also as inductive rules that resolve word lists to sentences. The speech recognizer reports a list of terminal symbols that are resolved to a sentence. If we trace the resolution again, we can get intermediate symbols, namely `direction, action`, etc., which are used to represent meanings of word groups.

Therefore, we programmed the Lisp parser to employ the same grammar for generating semantics. This eliminated a duplicated definition of the grammar, since a single grammar representation could be used both for recognition and form semantics analysis. Since the grammar is relatively simple and a list of word tokens given to the parser has already been qualified as a sentence, the parsing finishes almost in real-time. The Lisp parser produces the semantics of the following form:

```
(greeting hello)
(imperative    :action turn
               :direction right)
(imperative    :action goto
               :destination matsui
               :via corner)
(interrogative :person matsui
               :question location)
```

Dialogue States

Jijo-2's dialogue manager traverses between several dialogue states. The state transition diagram is depicted in Figure 5. The states are used to restrict the possible utterances to which *Jijo-2* reacts. For example, to begin a conversation, a user must say "hello" before any other statements. In the confirmation state, *Jijo-2* only listens to "yes" or "no".

This state transition network is needed to eliminate spurious utterances that are often generated as noises in an office. For example, we do not want an idle robot to react suddenly in response to people's occasional laughter near the robot.

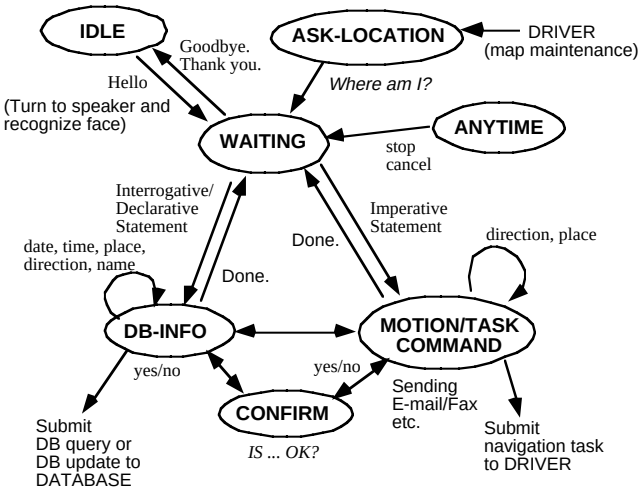

Figure 5: Dialogue states and transitions.

This state transition is also used to choose the most appropriate dictionary and grammar for the next speech recognition.

Commands like *stop* and *cancel* are recognized in any state. Obviously, *stop* should have the highest priority in any dialogue for safety reasons. As Bernsen (Bernsen, Dybkjoer, and Dybkjoer 1997) suggests, every practical spoken dialogue system should provide meta-commands to control conversation itself to make it robust and natural.

Dialogue Contexts

Each utterance of a human user only gives a fraction of information. For example, although "Turn" is a complete imperative sentence, the robot does not know which way to turn. In order to keep track of a series of relevant utterances and to construct semantics for a robot's behavior, we use the *dialogue context*.

Currently, we define five contexts: *query-context*, *update-context*, *identification-context*, *navigation-context*, and *call-context*. A context is defined to hold a number of required info and optional info as property variables of a *EusLisp* object. For example, a *query-context* that is created when an interrogative sentence is heard requires *person* property, and has *location*, *start-time*, and *end-time* as optional properties.

A conversation is guided to fulfill the required property by giving appropriate questions and confirmations. Optional properties may either be given in the utterance, assumed from the attentional state as described in the next section, or assumed by predefined default. For example, "a business trip" assumes the destination to be "Tokyo" unless it is explicitly spoken.

The state transition network is programmed in Prolog implemented in EusLisp.

Slot Filling by Managing Attentional States

A semantic representation of each utterance is given to the *dialogue manager* (Figure 6). The dialogue manager maintains an attentional state which indicates the relative salience of discourse referents (the individuals, objects, events, etc). Currently the attentional state is implemented as a total ordering. The function of the dialogue manager is to exploit the attentional state in order to accomplish (1) processing of under-specified input (information "unpackaging") and (2) natural-sounding language generation (information "packaging"). Though we have already shown the mechanism to control attentional states in *Jijo-2* (Fry, Asoh, and Matsui 1998), key points are summarized in the following subsections.

Japanese zero pronouns. In Japanese the problem of under-specified input is acute because in general the subject, object or any other argument to the verb is omitted from an utterance whenever it is clear from context. For example, the word "todokete" by itself is a well-formed Japanese sentence, even though its English translation "deliver!" is not. The missing arguments in Japanese are referred to as *zero pronouns* because of the analogy to the use of anaphoric pronouns in English. Thus the Japanese request *todokete*

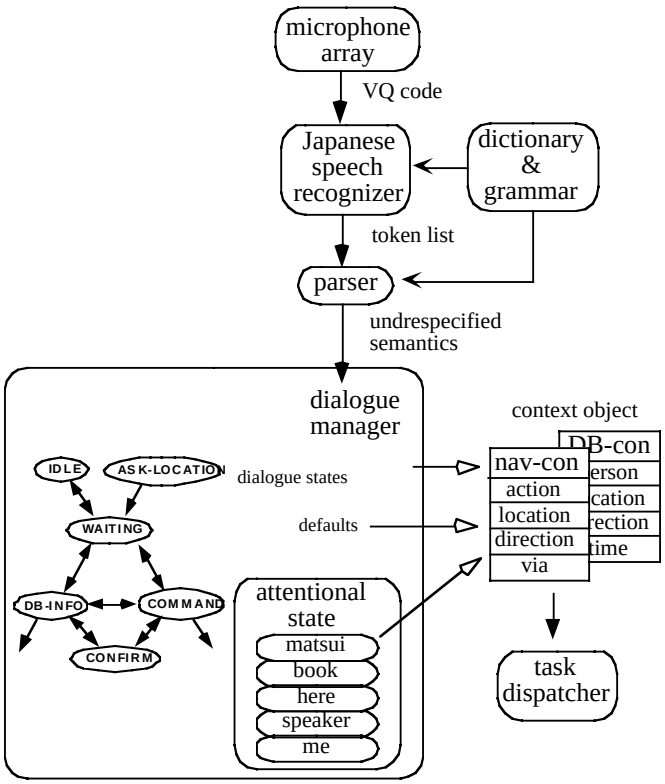

Figure 6: Flow of speech input processing.

corresponds roughly to the English request *(won't you) deliver this (to him)*.

Centering. To attack the zero pronoun ambiguity problem in a principled way we turned to the technique of *centering* (Grosz, Joshi, and Weinstein 1995). Centering is used for modeling discourse and for predicting the antecedents of pronouns based on the principle that the more salient an entity is in the discourse, the more likely it is to be pronominalized by a speaker. The centering model is broadly language-independent. The main language-dependent parameter is the criteria for ranking the salience of discourse referents.

For Japanese, we adopted the saliency ranking proposed by Walker et al. (Walker, Iida, and Cote 1994), given in Table 1. This ranking is based mainly on syntactic function, which in Japanese is indicated explicitly by post-positional particles such as *wa* and *ga* rather than by word-order position as in English. This ranking also takes into account empathy, or the perspective from which a speaker describes an event, which is an important feature of Japanese.

Language generation. In order to achieve the kind of natural-sounding dialogue, *Jijo-2* should also be able to *generate* speech output that takes into account the current dialogue context. The robot will sound awkward if it explicitly repeats the topic and subject of each sentence, or if it fails to use intonational stress to highlight newly-introduced discourse entities. For example, the highest-ranking entity in the attentional state usually does not need to be mentioned explicitly once it is established, so *Jijo-2* omits it from its spoken output.

1	Topic (marked by *wa*)
2	Empathy (marked by certain verbs)
3	Subject (marked by *ga*)
4	Indirect Object (marked by *ni*)
5	Direct Object (marked by *wo*)
6	Others (adjuncts, etc.)

Table 1: Predicted salience of entities

Integration with Database, Navigation, and Vision

Jijo-2 is an integrated robot that makes dialogue to provide services combined with its navigation and sensing capabilities. When all required slots in a context object are filled and confirmation is taken, the dialogue module tries to execute a task making use of these capabilities.

Database Behaviors

If the current context is a *query-context* or *update-context*, the task is dispatched to the *database* module. Normally, a response is immediate and the dialoguer can pronounce the result for the query or update task. If it takes long, the user may want to start another conversation, which is properly handled by the dialogue module, since the modules are all running concurrently in an event-driven manner. The database about people's schedule and location is implemented on a *postgres* server, which also provides web-based access. Therefore, once a dialogue about a user's schedule updates the database, the information is made available to public access.

Navigation Dialogue

For *navigation-* or *call-context*, the task is dispatched to the *driver* module. The map is maintained in the *driver* module, although it is learned through dialogue. Therefore, a situation where the dialogue module commands the driver to go to someone's office, but the driver does not know how to reach there, can happen. In this case, the *driver* module requests the *dialogue* module to ask for navigation instructions. During a navigation, the *driver* might encounter an unexpected landmark, which is usually a *close-to-open* event (an open space is found) from the sonar. This also leads the dialogue module to conduct a conversation to confirm the location.

Jijo-2 can continue dialogue while it navigates in a corridor. If the dialogue can be handled within the module or with the database like a query about the current date and time, it is locally processed without interfering with the navigation. But, of course, if the dialogue contains commands to stop navigation and to change destination, the dialogue module retracts the current command to the driver and restarts another behavior.

Face Recognition

Currently, the *Jijo-2's* vision system is used to look for a human user and to identify the person. When *Jijo-2* hears "hello" while it is in the waiting state, it turns to the sound source direction, and invokes the skin color detector. Moving the pan-tilter, the vision module tries to locate a human face at the center of the view. Then the face recognizer module is invoked.

The face recognizer is based upon higher order local correlation (Kurita et al. 1998). The vision module memorizes a face as a feature vector of 105 dimensions after taking at least as many shots of training images. For a particular sample image, it can tell the most plausible person name by computing distances to preobtained feature vectors in the discriminant analysis space. If the recognition succeeds, the vision module can provide the person's name to be inserted as the speaker's name in the attentional state stack of the dialogue manager.

Example Dialogue and Behavior

Figure 7 illustrates two patterns of dialogue between human users and *Jijo-2* involving several different behaviors.

Dialogue (a) is composed of simple motion commands. *Jijo-2* rotates its body by the *turn* command, and pans the camera by the *look-at* command. Though direction keyword is required for both, *Jijo-2* can assume it simply from the previous utterance.

Dialogue (b) begins with "hello", which makes *Jijo-2* turn to the person's direction. Then the robot captures the face image to recognize the person. If the recognition succeeds, the robot can know the speaker's name, which is then used to exclude misunderstandings in dialogue such as a command to send email to the speaker.

When the user asks *Asoh-san wa doko* (*Where is Mr. Asoh?*), Mr. Asoh is established as the topic of the dialogue, and therefore as the most salient entity in the attentional state. The answer given by *Jijo-2* is the under-specified response *Honkan ni imasu* (*(he) is in the main building*). Because it pronominalizes the current topic, this response is smooth and natural-sounding Japanese.

In the next sentence, the subject and indirect object of the verb *yonde* are omitted by the speaker in favor of zero pronouns. In this case, the most salient antecedent, which is Mr. Asoh, is found as the suitable referent. The *call-context* proposes sending an email message to Mr. Asoh. In the following sentences, other under-specified utterances are given appropriate antecedents or default values.

This inference for referents is not only useful to make Japanese dialogue natural and speedy, but also to attain the better speech recognition.

Performance Evaluation

Evaluating the performance and user-friendliness of integrated systems combined with human factors like Jijo-2 is a difficult problem. Except for the microphone array system,

H: Migi wo muite. (turn to the right)
R: *turn to the right*
H: Hidari. ([turn to the] left)
R: *turn to the left*
H: Ushiro. ([turn to the] back)
H: Hidari wo mite. (Look at the left).
R: *Look at the left*
H: Migi. (Right)
R: *Look at the right*

(a) Simple motion commands: the same direction keywords can cause different behaviors.

H1(Matsui): Kon-nichiwa. (Hello.)
R: *turn to the sound's direction, locate the person's face, and recognize.*
Konnichiwa, Matsui-san. (Hello, Matsui.)
H1: Asoh-san wa doko? (Where is Asoh?)
R: Honkan ni imasu. ([He is] in the main building.)
H1: Yonde kudasai. (Please call [him].)
R: Asoh-san ni me-ru wo okurimashouka?
(Shall I send E-mail to him?)
H1: Hai. (Yes.)
H1: Motomura-san wa ? ([Where is] Motomura ?)
R: Office ni imasu. ([He is] in [his] office.)
H1: Tsureteitte kudasai. (Please take me [to his place].)
R: Motomura-san no heya ni yukimasu.
([I'll] go to Motomura's office.)
plan the path to the office, and execute the plan
R: Motomura-san no tokoro ni tsukimashita.
([We are] at Motomura's office.)
H2(Motomrua): Shokuji ni demasu.
([I'm] going out for lunch.)
R: Motomura-san wa shokuji de yoroshii desu ka?
(Is Motomura going out for lunch ?)
H2: Hai. (Yes.)
R: Schedule wo toroku shimasu.
([I'll] register [the data] to the database.)
H1: Matsui mo. (Matsui, too.)
R: Matsui-san wa shokuji de yoroshii desu ka?
(Is Matsui going out for lunch ?)
H1: Hai. (Yes.)
R: schedule wo toroku shimasu.
([I'll] registrate [the data] to the database.)

(b) Zero pronoun filling in navigation and database tasks. Words in brackets are assumed.

Figure 7: Sample office dialogues

quantitative evaluation has not thoroughly been done because of the difficulty of controlling environment for fair comparisons. Here we just describe a qualitative evaluation from our demonstrations. Detailed quantitative evaluations should be done in the future.

Since we began building *Jijo-2*, we have demonstrated it more than 50 times to various visitors (Matsui and Jijo-2 Group 1999). In each demonstration, we made similar conversations as shown in Figure 7, and were guided from an office to another by the robot. Roughly speaking, 90% of demonstrations have been successful. During a demo, 10 to 20% of utterances were misunderstood in average. However, because possible actions taken in a particular dialogue state is limited and a confirmation is needed for a big task, the system did not go into a disaster. When the robot made a mistake, we could safely restart behaviors by canceling the improper actions and repeating a desired command.

The microphone array is proven effective to eliminate ambient noises at ordinary level. The speech recognition is confused mostly by unclear pronunciation or big noise, for example, loud laughter of visitors.

Related Work

Dialogue with a robot is a very old topic of artificial intelligence, going back to Nilsson's classic SHAKEY robot in the 1960s (Nilsson 1969). The influential SHRDLU system (Winograd 1972) also took the form of a dialogue with a robot. SHRDLU relied on syntactic analysis and was able to overcome ambiguity and understand pronoun references. Its success was due in large part to its restricted domain in the blocks world. More recently, Torrance investigated natural communication with a mobile robot (Torrance 1994), and Hashimoto et al. developed humanoid robots with dialogue capability (Hashimoto et al. 1997). Shibata et al. (Shibata et al. 1997) resolve ambiguities in route descriptions in natural language, using spatial relationships between the elements of the route. Horswill(Horswill 1995) integrated a simplified SHRDLU-like language processing system with a real-time vision system. These modern systems use language as a tool for interacting with the robot but are not particularly concerned with the principled application of linguistic or discourse theories.

Takeda et al. (Takeda et al. 1997) are realizing a dialog system for a mobile robot based on structured ontological knowledge and standard multi-agent architecture. In addition to dialog with a real robot, dialog with a simulated CG agent are also investigated recently (Hasegawa et al. 1993; Nagao and Takeuchi 1992).

Outstanding points of our system compared with these previous works are integration with mobile robot behaviors and a database to provide actual office services, and mechanisms to improve the robustness and quality of natural language spoken dialogue in real office situations.

Conclusion and Future Work

In order for a mobile robot to achieve natural Japanese conversation covering wider office task repertoire, we introduced four elemental techniques: 1) a microphone array, 2) multiple dictionaries, 3) dialog contexts, and 4) attentional states. 1) and 2) were proven to be effective for making speech recognition robuster in noisy environment. 3) and 4) enabled to form semantic frames for several kinds of tasks from fragmented utterances. The latter two also contributed to facilitate speech recognition and to make the dialog more natural for Japanese.

Based on these techniques we could successfully implement route guidance and database query tasks as well as

dialogue-driven topological map learning. To build a truly intelligent artificial creature, one cannot rely on a single information source. By integrating the sound source detector and the face recognizer with the dialogue module, we were able to achieve a better user-friendliness (Matsui and Jijo-2 Group 1999).

An important future goal is dialogue conversed with more than one user. In such a situation, we cannot naively rely on word symbols obtained from the recognizer. Rather, we will have to incorporate capabilities to distinguish speakers using visual cues, directions of the voice, and even the difference of the voice. Such capabilities will be crucial in the next-generation office dialogue systems.

Acknowledgments

The research and development of *Jijo-2* at the RWI-Center is jointly supported by Electrotechnical Laboratory (ETL) and the Real World Computing Partnership (RWCP) of MITI, Japan.

References

Asoh, H., Motomura, Y., Matsui, T., Hayamizu, S., and Hara, I. 1996. Combining probabilistic map and dialogue for robust life-long office navigation. In *Proceedings of the 1996 IEEE/RSJ International Conference on Intelligent Rbots and Systems*, 807-812.

Bernsen, N. O., Dybkjoer, H., and Dybkjoer, L. 1997. What should your speech system say, *IEEE COMPUTER*, 30(12):25-30.

Fry, J., Asoh, H., and Matsui, T. 1998. Natural dialogue with the *Jijo-2* office robot. In *Proceedings of the 1998 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1278-1283.

Grosz, B., Joshi, A., and Weinstein, S. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2): 203-225.

Hasegawa, O., Itou, K., Kurita, T., Hayamizu, S., Tanaka, K., Yamamoto, K., and Otsu, N. 1993. Active agent oriented multimodal interface system, In *Proceedings of the Fourteenth International Conference on Artificial Intelligence*, 82-87.

Hashimoto, S., et al. 1997. Humanoid Robot -Development of an Information Assistant Robot Hadaly, In Proceedings of the 6th IEEE International Workshop on Robot and Human Communication, RO-MAN'97.

Horswill, I. 1995. Integrating Vision and Natural Language without Central Models, In *Proceedings of the AAAI Fall Symposium on Embodied Language and Action*, Cambridge.

Itou, K., Hayamizu, S., Tanaka, K., and Tanaka, H. 1993. System design, data collection and evaluation of a speech dialogue system. *IEICE Transactions on Information and Systems*, E76-D:121-127.

Johnson, D. H. and Dugeon, D. E. 1993. *Array Signal Processing*, Prentice Hall, Englewood Clifs, NJ.

Kurita, T., et al. 1998. Scale and rotation invariant recognition method using higher-order local autocorrelation features of log-polar images, In *Proceedings of the third Asian Conference on Computer Vision*, Vol.II, 89-96.

Matsui, T., Asoh, H., and Hara, I. 1997. An event-driven architecture for controlling behaviors of the office conversant mobile robot *Jijo-2*. In *Proceedings of the1997 IEEE International Conference on Robotics and Automation*, 3367-3371.

Matsui, T., and Hara, I. 1995. *EusLisp Reference Manual Ver.8.00*, Technical Report, ETL-TR-95-2, Electro-technical Laboratory.

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

Nagao, K., Takeuchi, A. 1994. Social interaction: multimodal conversation with social agents, In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 22-28.

Nilsson, N. 1969. A mobile automaton: an application of artificial intelligence techniques. In *Proceedings of the First International Joint Conference on Artificial Intelligence*, 509-520.

Shibata, F., Ashida, M., Kakusho, K., and Kitahashi, T. 1997. Communication of a symbolic route description based on landmarks between a human and a robot. In *Proceedings of the 11th Annual Conference of Japanese Society for Artificiall Intelligence*, 429-432 (in Japanese).

Takeda, H., Kobayashi, N., Matsubara, Y., and Nishida, T. 1997. Towards ubiquitous human-robot interaction. In *Working Notes for IJCAI-97 Workshop on Intelligent Multimodal Systems*, 1-8.

Torrance, M. 1994. *Natural Communication with Robots*, Masters Thesis, Massatusetts Institute of Technology.

Walker, M., Iida, M., and Cote, S. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193-233.

Winograd, T. 1972. *Understanding Natural Language*, Academic Press.