

Integration of Real-world Interaction Functions on the Jijo-2 Office Robot

Toshihiro MATSUI, Hideki ASOH, Futoshi ASANO
Takio KURITA, Isao HARA, Yoichi MOTOMURA
Katsunobu ITOU, John FRY

In order for a mobile robot to provide information services in real offices, the robot has to maintain the map of the office. Rather than a completely autonomous approach, we chose to interact with office people to learn and update the topological map using spoken dialogue. To successfully apply a speech recognition technology to conversation understandings in real offices, we implemented a multiple microphone array system and a context and attentional manager in the robot. The robot could demonstrate simple map learning, route guidance, and people's location information service.

§1 Introduction

In 80's, robots became common workforces in factories. In the next decade, the robot technologies found a challenging exploration task on Mars. Recent home users are excited by the idea of petting robot for entertainment and welfare. But in the middle of factories and homes, we believe there is a demand for office robots that can work as secretaries. At RWI Center, ETL, we have been developing a prototype office robot, called Jijo-2. We suppose office robots should be able to provide information services combined with locomotion, such as route guidance, finding people, delivery, schedule arrangement, etc.

Unlike factory robots for unmanned operations in arranged working environment, office robots are expected to have capabilities to autonomously adapt to environments and to communicate with people in natural manners. Therefore, the research topics are map learning, localization in the map, speech dialogue system, human recognition and the integration architecture.

From the viewpoint of machine control, programs for 2D motions of mobile robots are easier than the ones for manipulators with 6-DOFs, and there are already

practical mobile robot systems for delivery tasks in factories. These robots, however, use preprogrammed map or signal-emitting wires embedded in the floor for guidance. Whether a mobile robot can navigate in an unknown environment without preprogrammed information or artificial settings is a research issue^{1,7)}.

Next, for friendly human-robot interaction, spoken language interface plays an important role. Although the spoken language interface has long been regarded as the most important channel for human-robot interaction since the first robot was dreamt, no actual robot has equipped it. Speech recognition and dialogue techniques that have resulted from long AI research are now available. Our interest is in what will happen when we apply this AI technology to robots in real offices^{5,9)}.

A factory robot is expected to repeat programmed motion sequences in a predefined set-up. An office robot has to plan appropriate actions taking into account different modes of sensory information such as voice, sonar, images, and so on. A remarkable difference is in a selection of use of uncertain information from various sensors whereas factory robots can rely on crisp data and timing. Office robots requires an architecture that allows such dynamic behaviors²⁾. In the following

sections, the architecture, the map representation and the dialogue systems to handle real world situations are described in this order.

§2 Robot Architecture

For the basis of the mobile office robot, we use a mobile robot platform, Nomad-200 (Fig.1), of Nomadic Technologies, U. S. A. It has a set of sonar sensors for measuring ranges up to several meters in 16 directions, infrared proximity sensors for faster but inaccurate distance measurement in as many directions up to half a meter, and bumper sensors for collision detection. We added two TV cameras, a microphone array and a speech synthesizer. Inside the robot is a Linux computer capable of communicating with hosts via a radio ethernet.

As shown in Fig.2, the control software is structured in two layers consisting of reactive modules on the robot and deliberative modules for planning and dialogue implemented on the host. We call the former the reactive layer, and the latter the integrator layer. Since all function module are running in separate processes, con-

current behaviors as the robot can talk while it is navigating in a corridor can happen at one moment (Fig.1). The whole structure is a kind of a multi-agent architecture that allows dynamic addition, removal and replacement



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

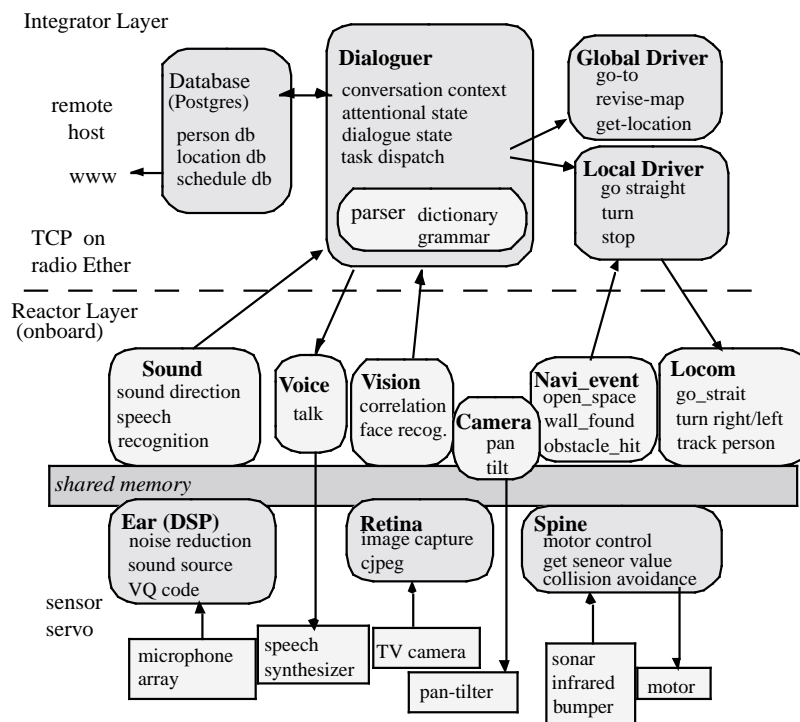


Fig.2 Organization of software modules of Jijo-2.

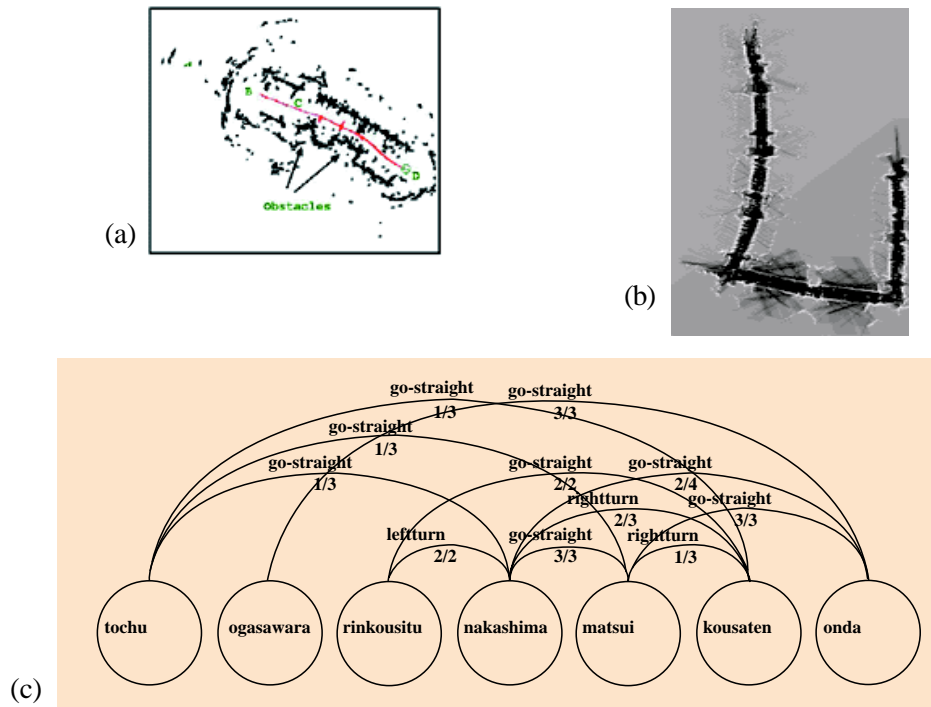


Fig.3 Map representations.
 (a) sonar profile in a corridor
 (b) sonar profile accumulated in a long navigation
 (c) topological map connecting location names with behaviors

of modules to adapt to different robot configurations and behaviors and to facilitate debugging ²⁾.

Each module is communicating through TCP sockets. Modules in the integrator layer request reactor modules to monitor events, i.e., *low voltage of batteries, finding an open space, loud sound*, etc. When a reactor module encounters an occurrence of the specified event, a reply message is sent to the requester, which invokes preregistered event processing function. Therefore, an integrator module usually consists of multiple event handling functions, each of which describes actions taken at the occurrence of an event and its chaining to other event monitoring. In this sense, there is no main program in the system.

§3 Navigation and Map

3.1 Navigation using topological maps

Jijo-2 mostly uses the sonar sensor for navigation. Accumulated sonar responses draw a distance profile as shown in **Fig.3** (a). Because of errors and ghosts, we

barely see big objects like a bookshelf and a trash basket. Further accumulation of sonar patterns draw a map like Fig.3 (b). Straight corridors are recognized as bent because of errors from the odometry sensors. It is not an easy task to localize a robot only with sonar information and the robot cannot solely rely on the odometer for long distance navigation.

The most basic navigation behavior of *Jijo-2* is to follow the corridor along walls. Sonars can produce range information accurate enough to direct steering to an open space avoiding obstacles. Also they can count the number of openings to office entrances if there is a change of distances to walls at these points. Therefore, the *Jijo-2* robot exploits a topological map which represents connectivity between these office entrances for the representation of its navigating environment. Each arc describes actions to reach the destination node together with a statistic information about the success and failure of the actions. This information is looked up and updated in the dialogue-guided map learning process described in the next section.

3.2 Dialogue-Guided Map Learning

We do not assume *Jijo-2* is given a map of an office environment in advance. From the moment the robot is put in an unknown environment, it begins to gradually learn the map from ordinary office workers. This is like a situation where a freshman to a new office is taken to important places and gets instructions about the location names, such as "Directors Office", "Meeting Room", etc. If he encounters with someone on the way, he is also told the name of the person by the instructor. *Jijo-2* does the similar task⁵⁾.

Since a map is represented by a network consisting of nodes for location names and of arcs for actions to traverse, the map learning is a task to create new location nodes, to record actions, and to update success/failure statistics. **Fig.4** depicts a map learning process of four locations starting from 1 through 2,3, and 4. As the result, nodes and arcs in Fig.3 (c) are added and updated. At 5th location, the robot was told to go to Matsui's office, and he began a navigation to the cor-

rect direction. However, he missed to recognize one of the doorways and reached 6th location. Because the distance he came was longer than his prediction, *Jijo-2* became uncertain about his location. *Jijo-2* asked his location, and replanned to go to the 7th location, where he could answer his location correctly. In this way, the statistic information attached to arcs is used to assess the reliability of his actions^{1,7)}.

§4 Spoken Language Dialogue System

4.1 Speech recognition

The most basic medium for interaction between people is spoken language conversation. *Jijo-2* equips with a speaker-independent continuous Japanese speech recognizer, which is a result of a research at ETL.

Today's commercial speech recognizers owe some of their nice performance to microphones set close to mouths. This microphone setting is important basically in two points: they are free from noises from the environment, and frequency spectrum does not differ much from the phoneme models which are also taken on-mic setting. If we apply the speech recognizer to voices at a distance (about 1 m), chances of misrecognition increase. Also, since most commercial systems collected correlation data between phonemes and words from big text corpus, they are good at recognizing readouts of texts like newspaper articles, but not very suitable for recognition of daily conversation.

To cope with the microphone problem, we developed a multiple microphone system for *Jijo-2*. **Fig.5** depicts the total structure of our speech recognition and understanding system. On the edge of the robot's top tray are eight omnidirectional microphones forming a half circle of 40 cm diameter. Based on the time difference for sounds to arrive at these microphones, the direction of the sound source is computed, and a focus beam is formed. By this technique we could upgrade the S/N ratio for consonants by 10 dB. This beam-forming processing is done by the first DSP (TI-C44/40MHz). The noise free voice signal is sent to the second DSP, which generates the vector quantization code every 10 ms.

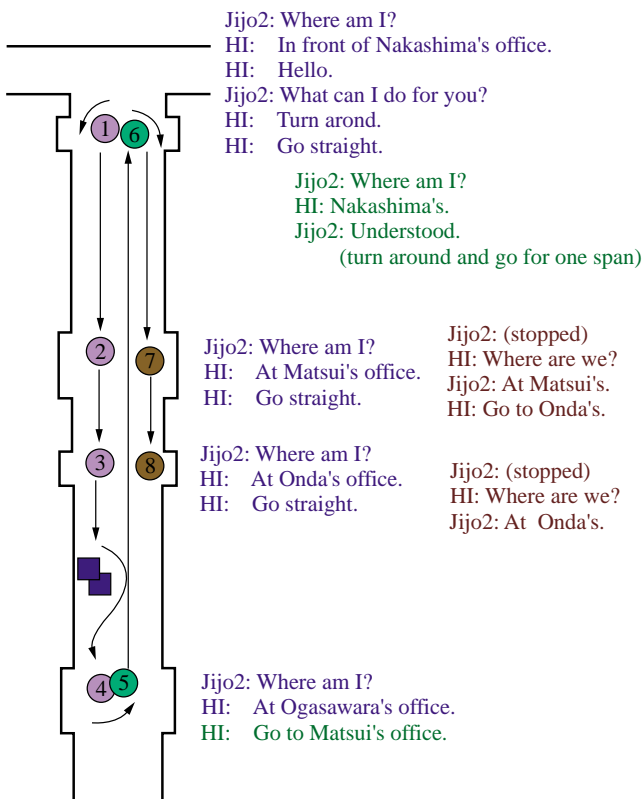


Fig.4 Dialogue-based map learning. Registration of locations and planning to a goal.

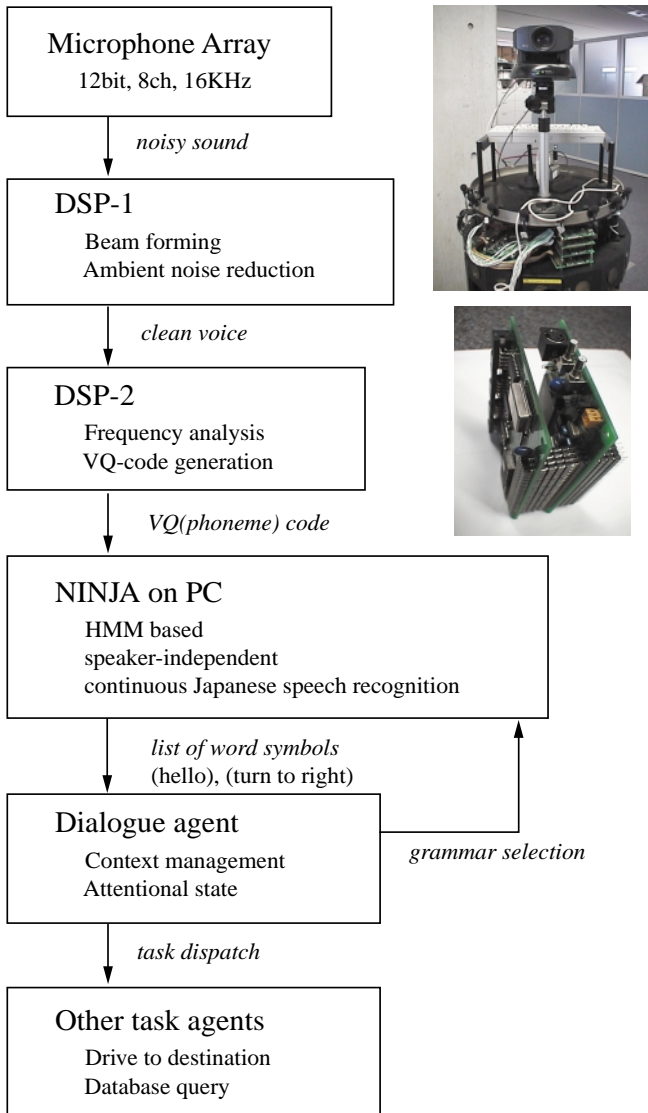


Fig.5 Flow of speech input processing.

Each VQ-code is an integer representing a phoneme in a short time frame. The VQ-code is sent to the PC and recognized by the speaker independent continuous Japanese speech recognizer, NINJA¹¹⁾. NINJA looks up word dictionary that contains about 150 words and a grammar for Jijo-2 tasks that contains about 300 rules.

4.2 Natural Japanese dialogue

The Japanese language has an obvious tendency to omit subjects, objects or any words if they are easily inferred from the context. Omission occurs more frequently in daily conversation than written texts. This makes the speech understanding complicated on the one hand, and makes the speech recognition and syntactic

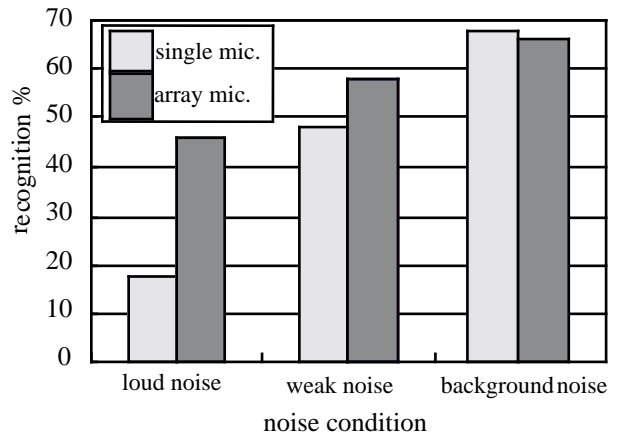


Fig.6 Speech recognition performance of multi-microphone array in various noise conditions.

analysis easier, since each utterance becomes shorter. Jijo-2 conducts a speech dialogue processing that exploits this characteristics of Japanese.

Jijo-2's dialogue manager traverses between several dialogue states. The state transition diagram is depicted in Fig.7. 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 voice inputs that are often generated as noises in an office, for example, occasional laughter near the robot. 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 for safety reasons.

4.3 Dialogue contexts

Each utterance of a user only gives a fraction of information. For example, though "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 7 contexts: *query-context*, *update-context*, *identification-context*, *navigation-context*, *call-context*, etc. A context is defined as to hold a num-

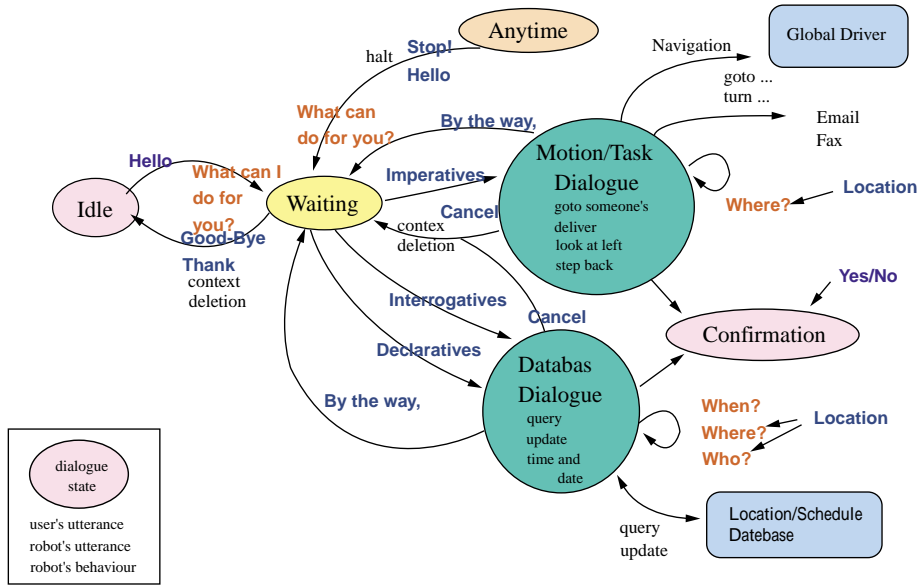


Fig.7 Dialogue states and transitions.

ber 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.

Fig.8 illustrates five dialogue contexts created in a series of utterances. 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 pre-defined 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.

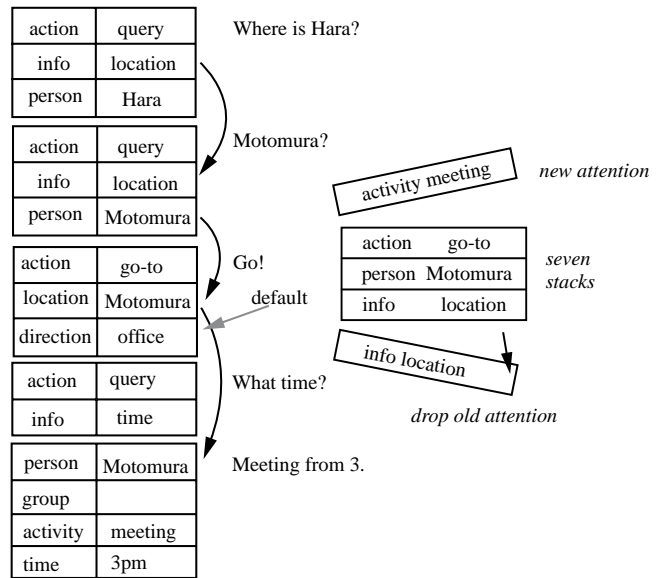


Fig.8 Dialogue context (a) and attentional state (b). Attentional state stack carries focus from one context to another.

4.4 Slot filling by managing attentional states

The dialogue agent 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”).

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* 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 centering¹²⁾. 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 referents⁴⁾.

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.

Attentional states managed by file cards

The attentional state is managed by a stack of seven file cards. Salient topics like location names, person names, times, etc., are recorded in the file card as they appear in a discourse. If a particular topic is needed in the next discourse analysis, the topic is searched in the file cards from the top. If found, it is given as a default value to the required/optional slot. Since new topics are always pushed from the top of the card stack, the topic at the bottom, which is the oldest attention, is removed to keep the number of attentional states to seven. The information in the attentional states are kept during the switching from one discourse guided by a particular dialogue context to another. Thus, the attentional state stack plays as a bridge to convey information from one script to another.

§5 Task Execution

These dialogue processes are conducted by the *dialogue* module in the integrator layer. Once an action to take is determined in a dialogue, the dialogue agent issues a request to other integrator agents. If it is a navigation to someone's office, a *go-to* command is sent to the driver agent. If the driver agent could successfully plan a path to a destination, it immediately starts navigation. If not, the driver agent requests the dialogue agent to ask the way how to reach there. In this manner, integrator agents do not have fixed precedence over other agents. The basic strategy is that if a task is attainable in the agent, it is processed locally, and if not, the task is dispatched to others.

5.1 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 time, 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 PostgreSQL server, which also allows web-based access. Therefore, once a dialogue about a user's schedule updates the database, the information is made available to public access.

5.2 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.

5.3 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 invoked¹⁰⁾.

The face recognizer is based upon higher order local correlation³⁾. 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.

§6 Example Dialogue and Behavior

Fig. 9 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 un-

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.

Fig.9 Samples of office dialogue

der-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.

§7 Future Works

In the year of 2000 and 2001, we will continue the research in the following directions. For more reliable speech conversation, we will reconstruct phoneme models reflecting the characteristics of real offices. Variations of distances between a speaker and the recognizer will also be taken into account. For this, a method to separate surrounding sound sources and to locate their positions is being studied⁶⁾. For the robot to efficiently navigate in offices, we will develop a scheme to hierarchically integrate the topological map representations with geometric representations obtained from a laser range finder and a landmark based vision system⁸⁾. For Jijo-2 to be able to adapt to general office environments, a learning architecture that represents relationships among events and behaviors in probabilistic networks and allows graphical programming will be studied. The most recent information is available at <http://www.etl.go.jp/~7440>.

Acknowledgements

Authors express sincere gratitude to Dr. Nobuyuki Otsu, Director of RWI Center, ETL, and to Dr. Jun-ichi Shimada, Director of RWC Tsukuba Research Center,

for providing us the opportunity of the Jijo-2 robot research.

References

- 1) Asoh, H., Motomura, Y., Matsui, T., Hayamizu, S., and Hara, I., Combining probabilistic map and dialogue for robust life-long office navigation. *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 807-812, 1996.
- 2) Matsui, T., Asoh, H., and Hara, I., An event-driven architecture for controlling behaviors of the office conversant mobile robot Jijo-2. *Proc. of 1997 IEEE International Conference on Robotics and Automation*, pp. 3367-3371, 1997.
- 3) Kurita, T., et al, Scale and rotation invariant recognition method using higher-order local autocorrelation features of log-polar images, *Proc. of third Asian Conference on Computer Vision*, Vol.II, pp. 89-96, 1998.
- 4) Fry, J., Asoh, H., and Matsui, T., Natural Dialogue with the Jijo-2 Office Robot, *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vancouver, 1998.
- 5) Matsui, T., Asoh, H., Fry, J., et al, Integrated Natural Spoken Dialogue System of Jijo-2 Mobile Robot for Office Services, *Proc. of 16th National Conference on Artificial Intelligence*, pp. 621-627, July, 1999.
- 6) Asano, F., Asoh, H., and Matsui, T., Sound Source Localization and Signal Separation for Office Robot "Jijo-2", *Proc. of IEEE International Conference on Multisensor Fusion and Integration Systems*, MFI-99, August, 1999.
- 7) Asoh, H., Coping with uncertainty using interactions with humans, *Proc. of IJCAI-99 Workshop on Reasoning with Uncertainty in Robot Navigation (RUR-99)*, Stockholm, August, 1999.
- 8) Asoh, H. and Matsui, T., A unified framework of map learning with a hierarchy of probabilistic maps, *Proc. of Field and Service Robots (FSR-99)*, pp. 86-91, Pittsburgh, August, 1999.
- 9) Asoh, H., Matsui, T., Fry, J., Asano, F. and Hayamizu, S., A spoken dialog system for a mobile office robot, *Proc. of Eurospeech-99*, pp.1139-1142, Budapest, September, 1999.
- 10) Hara, I., et al, Communicative Functions to Support Human-Robot Cooperation, *Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-99)*, October, 1999.
- 11) Itou, K., Hayamizu, S., Tanaka, K., and Tanaka, H., System

design, data collection and evaluation of a speech dialogue system, IEICE Transactions on Information and Systems, E76-D, pp.121-127, 1993.

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

(Accepted May 12, 2000)