Proceedings of the 2005 IEEE
International Conference on Robotics and Automation
Barcelona, Spain, April 2005

Real World Speech Interaction with a Humanoid Robot on a Layered Robot Behavior Control Architecture

Kazumi Aoyama and Hideki Shimomura
Sony Intelligence Dynamics Laboratories, Inc.
Takanawa Muse Building 4F, 3-14-13, Higashigotanda, Shinagawa-ku,
Tokyo, 141-0022, Japan
{kazumi,simomura}@idl.sony.co.jp

Abstract— **This paper presents requirements and a software framework for a humanoid robot aimed at speech interaction in a real world environment. We believe that for effective speech interaction with humans in the real world, not only is speech understanding and response generation required, but also other behavioral competencies are of paramount importance. Nodding, filler insertion, face tracking, and reactions to environmental stimuli during interaction are examples of such behaviors. In this paper, we discuss the importance of those competencies, termed "Naturalness Support Behavior", and list the requirements for their implementation. A layered framework that satisfies the requirements is proposed. The system is integrated into the EGO architecture, which is the behavior control architecture for the Sony humanoid robot QRIO SDR-4XII. We also present preliminary experimental results of speech interaction based on the proposed ideas and conducted on QRIO.**

Index Terms— **human-robot interaction, speech interaction, natural behavior**

I. Introduction

We have seen extensive research on personal robots in recent years [1][2]. Considering the fact that a large part of human-human communications is conversational, a robot that 'lives' with humans in daily life should be capable of speech interaction.

A large amount of spoken dialogue research relates to speech understanding and response generation and thus focuses on the meaning or content of the dialogue[3]. In this paper, we use the term "Content level Speech interaction Behavior (CSB)" to describe the behaviors based on that approach. Other research focuses on how to make the dialogue smoother by incorporation of paralinguistic interaction, such as nodding or filler insertion[4][6]. We call this kind of behavior "Naturalness Support Behavior (NSB)", which are rhythmically related and mutually synchronized between talkers, and play an important role in human-human communication [6]. According to this categorization, it can be said that each behavior in NSB serves as a basic function for natural speech interaction and could support many of the CSB behaviors.

We believe that for effective speech interaction with humans in the real world, not only is speech understanding and response generation required, but other additional behavioral competencies are of paramount importance. Research on spoken dialogue system has not focused enough on such competencies. In other words, NSB on a robot should be extended to cover a broader behavioral range than is available in current systems. For example, some personal robots have the ability of tracking the interaction partner's face[5] and there is little doubt that this function in spoken dialogue makes the interaction more natural. Also, in order to support a natural interaction, the robot should react as much as possible to the environment similar to how humans do. For example, when a human hears a loud sound while talking with someone, their attention to the dialogue may be drawn away by that sound. If a robot responds to it as a human does, it makes the co-existence of the robot in our environment appear more natural to us. Since we don't notice these facts unless we're interacting with a robot in the real world, they are rarely focused on in the spoken dialogue research community. Generally speaking, NSB are simple behaviors, and quick response is required because reaction speed is critical to convey naturalness. On the other hand, CSB requires heavier computation, such as speech recognition, understanding, or generation. CSB and NSB should cooperate for overall natural interaction. Thus CSB should be able to control the activation of NSB, e.g. some behaviors in CSB may want to inhibit NSB components so as not to interfere with their operation. It is necessary that the framework provides a means to control conflicts as well as provide cooperation between NSB and CSB. For example, during a dialogue, face tracking, which is one of the NSB, should track the same person the dialogue module in CSB is talking to.

In order to satisfy those requirements, we present a two-layered framework that we call "layered structure for Real World Speech Interaction (RWSI)". The upper layer or "CSB layer", deals chiefly with speech understanding and response generation. The lower layer or "NSB layer", deals for example with nodding, filler insertion, face tracking and natural reactions against environmental stimuli. In order to achieve quick response for the NSB layer, these two layers are put in different CPU processes and allowed parallel execution. Moreover, modules in the NSB layer can be shared by different modules in the CSB layer. Information sharing mechanisms between the layers are also available.

The paper is organized as follows. In the next section we analyze CSB and NSB and then propose the "layered structure for RWSI". Section III presents a brief explanation of the EGO architecture. In section IV the implementation is described together with examples of the human-robot interaction. We conclude in section V.

0-7803-8914-X/05/$20.00 ©2005 IEEE.

II. Analysis of Real World Speech Interaction Behavior

A. Content level Speech interaction Behavior (CSB)

CSB principally addresses speech understanding and response generation. Usually, behavior modules at this level are described using a state transition model, with one module created for one dialogue task because each task has a different dialogue flow. Processes in CSB are often computationally heavy and are not suitable for quick response, as no response can be generated until the speech recognition process is completed. Also, if some type of information retrieval is necessary, then it may take additional time to generate the response and the user must wait awhile before any reaction occurs from the robot. If the robot does not have some form of parallel behavior execution to 'fill' such dead-times, then the interaction tends to be unnatural as the robot simply stands without doing anything. CSB alone is clearly inadequate to provide smooth and natural interaction.

B. Naturalness Support Behavior (NSB)

The following behaviors are listed as examples of NSB.

1) Nodding A well-known group of NSB is paralinguistic reactions, such as nodding or filler insertion. These have been investigated in spoken dialogue research, and their positive effects demonstrated. Nodding is a head move action during the user's speech or at the end of it. It is taken as a common signal of acceptance of user speech in the interaction.

2) Filler insertion Filler insertion involves the activity of filling blank time in speaking with a sound, word, or phrase (such as "Let me see") and so on. In normal human-human conversation, it is used for taking time for consideration without disturbing the flow of conversation. This same idea is applied in human-robot conversation. If speech recognition is carried out only with the robot's onboard hardware resources, it takes time until the recognition completes. This delay is a critical problem for ensuring smooth and natural speech interaction.

3) Face tracking Looking at the interaction partner's face during communication is also very important. Talking with concurrent face tracking shows the interaction partner that the robot is concentrating on the speaker.

4) Reaction to the environmental stimuli A robot needs to naturally react against unusual environmental stimuli, for example, a loud sound or a sudden change in the visual field of view. These types of unexpected stimuli must draw the interaction partner's attention from the robot to the direction of the stimuli source to confirm what has happened. At that moment, it is unnatural for the robot to neglect the stimuli and keep the interaction moving as if nothing happened. This kind of feature has not been mentioned in general terms previously, but we believe it is very important for a robot operating in the real world.

C. Analysis for implementation and execution

As mentioned earlier, NSB requires quick response. For instance, since nodding provides important feedback to the partner during the course of the interaction, its timing is critical. Reaction to the environmental stimuli also requires quick response so that the interaction partner can reasonably attend to the offending stimuli.

Another important point is the relationship between CSB and NSB. As mentioned before, at first, the functions of NSB seem independent from CSB and they can be shared across modules in CSB. However, the usage of NSB may be different by each module within CSB. For example, in normal dialogue, nodding or filler insertion is effective. Therefore, a CSB module may profit from the execution of almost all NSB modules. On the other hand, for example, when the robot is visually learning the partner's face, nodding may wrongly interfere with the learning process. If that case, even though nodding is useful during a dialogue sequence, the face learning module in CSB may prefer to temporarily stop the nodding NSB. This kind of control framework is necessary for implementing robust and flexible dialogues. Moreover, these two behavior groups need to have a mechanism for sharing information. For example, the person whom the CSB recognizes must be the same person to whom the NSB reacts.

In order to permit more natural human-robot interaction using the above mechanism, there is one additional requirement. While engaged in interaction, if the NSB implements a reaction to an environmental stimuli then the robot may, for example, look in the direction of a loud sound, thus interrupting the ongoing interaction. It is then required that the robot be able to return to the interrupted interaction, should it judge that the environmental stimuli does not need further attention. The framework must allow for a behavioral interruption-resume mechanism.

D. Layered structure for RWSI

In order to satisfy the requirements mentioned above, a two-layered framework for natural speech interaction is proposed that we call the "layered structure for RWSI". Fig.1 shows the framework. We define the role of the two layers as follows:

1) CSB layer This is the upper layer which contains the CSB modules. Various dialogue modules and their switching mechanisms are implemented here.

2) NSB layer This is the lower layer which consists of the NSB modules, such as face tracking, nodding, filler insertion, and natural reflexive responses to unusual environmental stimuli.

These two layers are assumed to be processed independently. From an implementation point of view, the two layers are located in two different processes. This separation allows the NSB layer to react quickly without influence of the CSB layer. When one of the NSB layer modules becomes active, the CSB layer may be interrupted. During this interruption, the CSB layer module stops, but it maintains its own status, and after the interruption, it then resumes the process.

Alternatively, the CSB layer module can control the activation of NSB layer modules. More precisely, each module in the CSB layer can define the set of usable

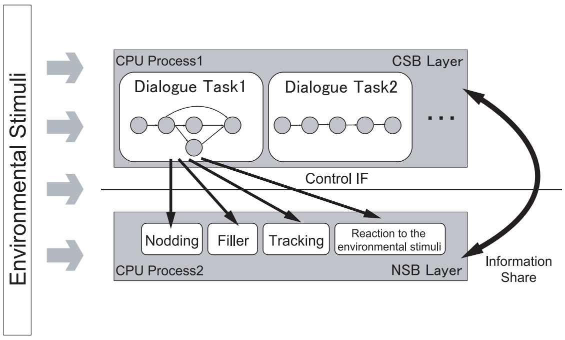

Fig. 1. Layered structure for RWSI

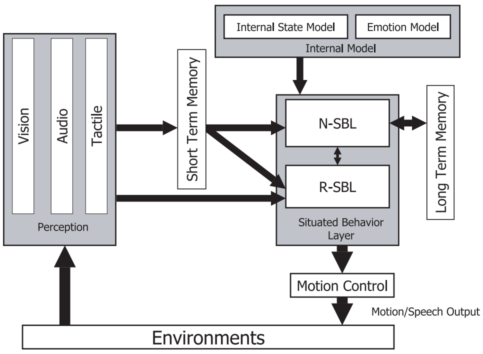

Fig. 2. Overview of the EGO Architecture

modules in the NSB layer. The framework also provides an interface and mechanism for sharing information between the two layers. In fact, information about the interaction partner needs to be shared between the two layers if they intend to cooperate, e.g., act on the same target. For example, a module in the CSB layer may set the target object for modules in the NSB layer, such as nodding or face tracking. The information sharing mechanism and the interface for proper coordination of the NSB and CSB layers play important roles for satisfying the requirements mentioned earlier.

III. Overview of EGO architecture

The EGO architecture (Emotionally GrOunded Architecture) is the behavior control architecture of the autonomous robot QRIO SDR-4XII[7] implemented by our development team. Fig.2 illustrates the structure of the EGO architecture. The basic idea arises from an ethological model[8]. During the design phase of the EGO architecture, a ”layered structure for RWSI” was taken into consideration. All the requirements for natural speech interaction discussed in section II have been satisfied. The EGO architecture treats not only speech interaction but also other types of autonomous behavior. In a sense, we can say that it covers a wider area than the ”layered structure for RWSI”. The behavior selection mechanism of EGO architecture depends principally on the ”homeostatic regulation rule” that evaluates both external stimuli and ongoing internal drives[10]. It is based on the determination of behavior selection values. The calculation is performed for all the behavior modules in its repertoire by referring to the internal state variables and external stimuli[9]. The robot selects and executes the behavior that has the highest behavior selection value. The behaviors implemented on QRIO include, for example, ”to search a ball and kick to increase the exercise internal value,” to take a rest to reduce the tiredness internal value” and so on.

In this architecture, speech interaction is treated as one of the robot's behaviors. We introduced a variable corresponding to the desire to talk to a human into the internal state model of the robot. A speech interaction behavior is occurs autonomously to regulate this value.

We now explain the essential modules of the EGO architecture involved in speech interaction. See[10] for descriptions of the other modules appearing in Fig.2.

1) Situated Behavior Layer Situated Behavior Layer (SBL) is the behavior control module for QRIO [11]. SBL is comprised of many behavior modules that are integrated in the form of a tree structure. This module houses the behavior selection mechanism mentioned before. SBL is designed so as to allow multiple behavior modules to be executed in parallel unless a robotic resource conflict occurs. Each behavior module is described with the state transition model. When a module is interrupted by another module, the interrupted module suspends its execution maintaining its own execution status, then after the interruption, it resumes. Moreover, the EGO architecture allows for multiple SBLs and defines protocols between the SBLs. SBL is equipped with two types of interfaces for inter-SBL connections. One is for sending and sharing information of robot's behavioral target. The other is for controlling connected SBLs. Activation of each module in an SBL can be controlled from another SBL through the interface. Therefore, a multi-layered behavior control architecture can be realized using the architecture.

2) Short Term Memory The EGO architecture includes Short Term Memory (STM). STM integrates the results of perception and recognition (face, sound direction, etc) considering their temporal-space relationship. It also maintains the results for a while. When a behavior is interrupted by another one, STM keeps and provides environmental information for later resuming the interrupted behavior.

3) Long Term Memory The EGO architecture also has a Long Term Memory (LTM). LTM consists of two modules: one is an associative memory using a neural network, and the other is a memory to keep each user's personal information acquired through speech interaction. The associative memory is used for memorizing a combination of the face identification result, speaker identification result and the user's name. This corresponds to memorization of a person newly encountered by the robot. Another memory stores the user's personal information gathered by questioning the user. The combination of these two forms of memory in LTM realizes a dialogue that can be based on an individual's personal topics.

IV. Experimental Implementation with EGO architecture

A. Correspondence of layered structure for RWSI to EGO architecture

We implemented the ”layered structure for RWSI” with two SBLs as shown in Fig.3. CSB is implemented in the upper layer named the Normal Situated Behavior Layer (N-SBL). NSB is located in the lower layer named the Reflexive Situated Behavior Layer (R-SBL). Note that although we focus on speech interaction related behavior, these two layers usually contain many other behaviors to serve several different purposes. Modules in each layer can access STM and LTM, and can output motion and speech commands. R-SBL is allowed to receive sensor information directly to achieve rapid response times, while N-SBL only receives processed information from STM. This is, of course, not a limitation of the architecture, but a design choice for our experiments. Each SBL selects and executes behaviors independently. We use the N-SBL/R-SBL connection protocol to share information between NSB and CSB and to control NSB from CSB.

B. Modules in NSB

The implemented NSB modules are nodding, filler insertion, face tracking, and reaction to loud sounds (one of the environmental stimuli). More details appear below.

1) Nodding Nodding can be fired when it receives the speech signal from the speech detector, or it finds that the current speech is longer than a threshold value. Within this function, nodding motions have some variation in neck angle and speed in order to avoid a monotonous reaction.

2) Filler The Filler module is activated when an interaction partner's speech ends but speech recognition needs some additional time to complete its processing. After receiving the speech end signal from the speech detector, it predicts the required processing time for speech recognition based on the length of the speech. Then it selects and outputs a corresponding length of filler phrase. The variation in the reaction is also very important in this case. Some phrases including ”null-phrase”, that intentionally say nothing, are provided as filler patterns.

3) Face tracking The Face tracking module in NSB receives information that designates the target face from one of the CSB modules. This means that CSB modules decide with whom the robot interacts. This target face information consists of two recognition results: the face identification result and skin color. When the tracking module cannot find the designated target face using a face identification process, but it believes it detects the existence of the partner using color recognition, the tracking module continues its process. Even if the target disappears from both face identification and color recognition processes, it tries to acquire the target again with formally designated face information and resume tracking if possible. In real world recognition, this kind of ”missing face” problem often happens. This retry mechanism increases robustness in face tracking. If the tracking behavior module is interrupted by another head-moving behavior such as nodding, it should resume on the same face. Since STM makes available the target face information with position, the robot can easily accomplish this task even if the target is out of the robot's view, as long as the target does not move extremely far from its previous location. In this case, the tracking behavior module turns the robot head towards the person with whom the robot was talking before, and resumes tracking her face.

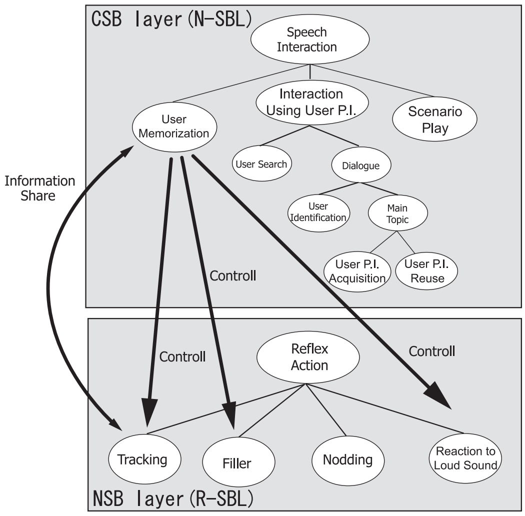

Fig. 3. Experimental Implementation of Speech Interaction Behavior

4) Reaction to loud sound This module is responsible for turning the robot's head toward a loud sound source. We implemented hand-clapping detection in the speech processing part as one of the sound stimuli because a clapping sound can be robustly recognized. When the robot recognizes the clapping sound, it turns toward that direction. The first objective of this behavior is to show its naturalness regarding the sound stimuli. However, we found that this can also be used to draw the robot's attention to a user when the robot has not previously noticed their existence. With this function, a user can assist the robot in finding her in a very natural way.

C. Modules in CSB

The CSB modules in this experimental implementation are: (1) user memorization and user identification module, (2) user personal information acquisition and reuse module, and (3) scenario play.

1) User memorization and user identification QRIO has an online new word acquisition function[12] and is equipped with speaker identification and face identification modules, both capable of online learning. Using these features combined with associative memory, QRIO is able to memorize a user profile through speech interaction. The user memorization dialogue module is implemented in CSB and it is used when QRIO meets a person for the first time, i.e., there exists no profile for her. The dialogue follows a predefined sequence and controls online learning. To memorize the user's face and voice, this module sends learning start commands to the face identification and speaker identification modules. When both modules have learned the features well enough, they send IDs for the newly acquired target (person) to the user memorization dialogue module. When the robot learns the new person's face and voice online, it engages the user in a dialogue that first asks for their name. After that, the robot memorizes the face and voice IDs with the new user name within the associative memory as a user profile. When the robot later meets that same person again, it can recognize her by their face or voice and recall her name. If it does not have enough confidence in the identification, the robot may ask her name to confirm it.

2) User personal information acquisition and reuse We also implemented a dialogue module that acquires a user's personal information and uses it during dialogue interaction[13]. When this dialogue module is activated, the robot asks the user her personal information and memorizes it in LTM. Then when the robot meets the same person later, it can engage in a dialogue whose content is based on the previously acquired information. Since the robot has to know with whom it is talking in order to perform correct dialogue, the user memorization and identification, explained above, is very important. Items used for personal information are designed and fixed in advance, but the order of questions or information reuse is variable. These items have been selected using the following criteria: (a) to make a user feel familiarity with the robot, and (b) provide variable information through time for reusability. For example, the person's favorite food or objects, birthday, etc., are included within the user's preferences.

3) Scenario play A development environment based on simple state machine transitions is used to implement a simple dialogue based on a scenario definition. For example, the robot's self-introduction or an explanation of the robot's features can be easily created. Of course, a scenario can contain conditional branches triggered by user response, but the currently provided function in this framework is limited. However, a robot interaction designer can easily implement a simple dialogue using the framework.

Since these three modules form a part of the EGO architecture, module selection is accomplished based on the ideas explained in Section 3.

D. *Interfaces between layers*

When a dialogue module in CSB is selected, it sends an activation control command to the NSB modules. In this way, the CSB module designates the NSB modules allowed to execute in parallel with it. When the dialogue module finds a target that should be shared with NSB, it provides this information to NSB. Since NSB modules also refer to STM, it can then execute its own actions against the same target, for example, when the tracking module in NSB tracks the same target as the one that a CSB module focuses on.

E. *Interruption-resume mechanism*

We place higher priority on NSB than CSB in execution for natural interaction. When an NSB module executes, if an active CSB module has a robotic resource conflict with the NSB module, the CSB module is forced to suspend its execution. When the NSB module finishes its action and the CSB module is now able to use the previously conflicted robotic resource, it continues execution using the interruption-resume mechanism that is supported by SBL.

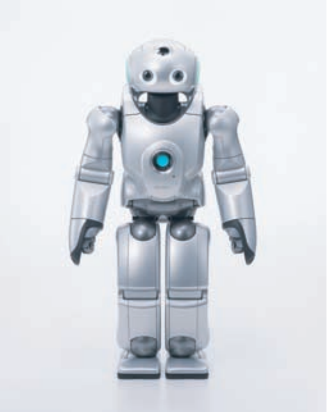

Fig. 4. SDR-4XII QRIO

F. *Overview of QRIO SDR-4XII*

An overview of our research platform, QRIO SDR-4XII, is now presented. QRIO is a biped robot with a height of approximately 60 cm and a weight of about 7 kg. It has 38 DOF as effectors, and is also equipped with LEDs and a speaker for talking. Fig.4 illustrates QRIO's appearance. Regarding sensing devices for interaction with people, it has touch sensors on its head and shoulders, and multi-microphones and stereo camera in its head. Regarding the software, face detection, face identification, and distance measurement using the stereo cameras are implemented. We realized obstacle avoidance for walking by QRIO by combining 3D recognition of the environment and motion control[14]. As for speech interaction related functions, sound source localization, speaker identification, grammar-based speech recognition, Large Vocabulary Conversational Speech Recognition (LVCSR) and Text To Speech (TTS) are all implemented.

G. *Experiment with QRIO*

Speech interaction experiments were performed using QRIO. QRIO's typical behavior realized with the proposed framework is described below.

When QRIO, finds an unknown person in the environment, it starts the user memorization dialogue to produce a user profile. When the user memorization dialogue module in CSB is selected, the corresponding modules can be activated in NSB, e.g., the tracking behavior, but not the nodding behavior because it can disturbs face image learning. Reaction to environment stimuli (e.g., loud sounds) is almost always selected in NSB. The user memorization dialogue module sends the acquired information of its interaction partner to the tracking behavior module to share it.

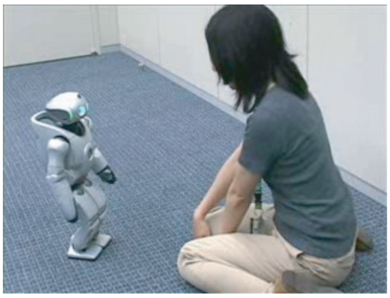

Fig. 5. Illustration of speech interaction with QRIO

After it has learned the user, if the speech interaction behavior value is still high, it would start the user personal information acquisition dialogue. In this case, all NSB modules can be selected. If the robot recognizes the clapping sound during this dialogue, it turns towards the sound source and may say, ”What happened? ” as a natural reaction created by its reaction from the loud sound behavior module in NSB. During this behavior, the user personal information acquisition dialogue module and tracking behavior module suspend their execution because they have a conflict of robotic resources with the NSB module. After completing its reaction to the sound, the suspended modules can then resume. A picture during a speech interaction between QRIO and a user appears in Fig.5. A transcription of a sample dialogue sequence is given below.

-QRIO meets an unknown user.
(QRIO is tracking the user's face.)
R(Robot): Please tell me your name.
(Face and speaker identification modules start learning.)
U(User): My name is Kazumi.
R: Ah-ha (filler), Kazumi, right?
U: Yes.
R: Please show me your face a little longer.
R: Thank you.
(Stop learning and memory to associative memory.)
R: May I ask you something?
U: Sure.
R: *(nodding)* Which food do you like?
(Someone claps hands at the right side. It turns to the sound source direction.)
R: Hey, what happened?
(It recognizes the clapping position but it prefers to talk with the user.)
R: Which food do you like?
U: I like apples.
R: *(nodding)* Umm *(filler)*, you like an apple, do you?
U: *(nodding)* Yes.
R: Thank you for teaching me.

V. Conclusion

In this paper, we proposed a ”layered structure for RWSI” and provided preliminary speech interaction experimental results for its implementation in the EGO architecture on QRIO. We classified robot dialogue behaviors into two layers: CSB and NSB. We modularized NSB and implemented it in the lower layer; CSB is implemented independently in the upper layer. This separation worked quite well and we confirmed the effectiveness of NSB in speech interactions. Furthermore, as a result of this design, the role of the two types of behavior groups are separated and clarified, and it allows robot interaction developers a pathway for an easier implementation.

We have not yet performed a solid quantitative evaluation, but our preliminary results from interviews with test users show that the robot exhibits a more natural overall behavior when using NSB. The reaction speed of NSB is especially accepted very positively by users. Although in this paper, we focused on speech interaction, but many other types of behavior can be implemented within QRIO. Some other examples are described in[9].

Acknowledgment

We greatly appreciate all researchers and engineers in Sony Entertainment Robot Company and especially the EGO architecture development team members. We would like to express our thanks to all members involved in QRIO development in Sony Corporation.

References

[1] S.Ohnaka, et al, ”The introduction of the personal robot PaPeRo,” IPSJ SIG Notes, Vol. 2001, No.68 pp.37-42,2001.

[2] H.Ishiguro, et al, ”Robovie: an interactive humanoid robot,” Industrial Robot, Vol. 29, No.6, pp.498-503, 2001.

[3] M.Nakano, et al, ”Handling Rich Turn-Taking in Spoken Dialogue Systems,” Proc. of Eurospeech99, pp.1167-1170, 1999.

[4] J.Hirasawa, et al, ”Implementation of coordinative nodding behavior on spoken dialogue systems,” Proc. of ICSLP, pp.2347-2350, 1998.

[5] Cecilia Laschi, et al, ”Visuo-motor Coordination of a Humanoid Robot Head with Human-like Vision in Face Tracking”, Proc. of ICRA 2003, pp.232-237, 2003.

[6] T.Watanabe, et al, ”InterActor: Speech Driven Embodied Interactive Actor,” Proc. of the 11th IEEE International Workshop on Robot and Human Interactive Communication, pp.430-435,2002.

[7] M.Fujita, et al, ”An Autonomous Robot that Eats Information via Interaction with Humans and Environments,” IEEE International Workshop on Robot and Human Interactive Communication, pp.383-389, 2001.

[8] Arkin,R.C., et al, ”Ethological Modeling and emotional basis for human-robot interaction,” Robotics and Autonomous System, vol.42, pp.191-201, 2003.

[9] T.Sawada, et al, ”Behavior selection and Motion modulation in Emotionally Grounded Architecture for QRIO SDR-4XII”, Proc. of IROS-04, pp.2514-2519, 2004.

[10] M.Fujita, et al, ”Autonomous Behavior Control Architecture of Entertainment Humanoid Robot SDR-4X,” Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.960-967, 2003.

[11] Y.Hoshino, et al,”Behavior description and control using behavior module for personal robot,” Proc. of ICRA 2004, pp.4165-4171, 2004.

[12] H.Shimomura, et al, ”Autonomous Entertainment Robot and Speech Dialogue”, The Japanese Society for AI, SIG-SLUD-A202, pp.21-16, 2002.

[13] K.Aoyama, et al, ”Spoken Dialogue on Robot Using Personal Information”, The Japanese Society for AI, SIG-SLUD-A301, pp.31-36, 2003.

[14] M.Fukuchi, et al,”Obstacle Avoidance and Path Planning for Humanoid Robots using Stereo Vision,” Proc. of ICRA 2004, pp.592-597, 2004.