AI & Soc (2007) 22:185–196
DOI 10.1007/s00146-007-0131-4

OPEN FORUM

A two-layered approach to communicative artifacts

Yong Xu · Tatsuya Hiramatsu · Kateryna Tarasenko · Toyoaki Nishida · Yoshiyasu Ogasawara · Takashi Tajima · Makoto Hatakeyama · Masashi Okamoto · Yukiko I. Nakano

Received: 31 March 2005 / Accepted: 29 August 2006 / Published online: 29 June 2007

Abstract A key issue in social intelligence design is the realization of artifacts that can fluently communicate with people. Thus, we proposed a two-layered approach to enhance a robot's capacity of involvement and engagement. The upper layer flexibly controls social interaction by dynamic Bayesian networks (DBN) representing social interaction patterns. The lower layer improves the robustness of the system by detecting rhythmic and repetitive gestures. We designed a listener robot that can follow and record humans' explanation on how to assemble and/or disassemble a bicycle. The implementation of this system is described by assembling the key algorithms presented in this paper.

Introduction

In recent years, the functions of artifacts such as electrical appliances, personal computers, and personal service robots have become increasingly powerful and

Y. Xu (✉) · T. Hiramatsu · K. Tarasenko · T. Nishida
Graduate School of Informatics, Kyoto University,
Yoshida-Honmachi, Sakyo-Ku, Kyoto 606-8501, Japan
e-mail: xuyong@ii.ist.i.kyoto-u.ac.jp

Y. Ogasawara · T. Tajima · M. Hatakeyama
Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan

M. Okamoto
Katayanagi Advanced Research Laboratories, Tokyo University of Technology, Tokyo, Japan

Y. I. Nakano
Department of Computer and Information Sciences,
Tokyo University of Agriculture and Technology, Tokyo, Japan

complex. In order to completely exploit the functions of such complex artifacts, we need to establish a means of communicating intentions between humans and these artifacts without increasing the burden to human users.

Watanabe and Ogawa (2001) developed an InterRobot that uses facial expressions, postures, and gestures to support remote communication between human beings. When a human user speaks to the InterRobot, it generates gestures or postures according to the expressions of the remote user, as if the user is having a face-to-face conversation with another person. ROBITA, developed by Matsusaka (Matsusaka et al. 2001), is a type of humanoid robot that can participate in a group conversation with two human users. It can use voices and gestures according to the users' behaviors. Unfortunately, the scope of the previous work was severely limited since little attention was paid to nonverbal communication behaviors in the situation where humans are required to share some objects as a referent while interacting with robots.

The main content of this paper comprises a theory on establishing user involvement by joint attention and a two-layered architecture of a human-robot interaction system. We are developing a listener robot that can follow and record a human speaker's explanation of certain procedures of the construction/assembly and/or dismantling/disassembly of a bicycle.

User involvement by joint attention

A robot should be able to follow the social norms of human society in order for it to naturally involve itself in communication with a human. In this paper, we address user involvement (Okamoto et al. 2004) as a key concept with regard to natural human-robot communication. User involvement occurs when humans willingly engage in or are forced to be involved in a virtual world wherein, computers simulate human-robot communication. We argue that joint attention plays a critical role in achieving user involvement in human-robot communication.

We consider two conditions that need to be satisfied in order to establish user involvement. First, cognitive/communicative reality should be achieved. The user should feel that the human-computer interaction is real. Second, two or more cognitive spaces should be linked. The user should smoothly move in and out of at least two cognitive spaces.

As Reeves and Nass (1996) pointed out, humans are likely to respond to artifacts as if they were humans. This implies that if the robot reacts to the user in unnatural ways, she/he may assume that the robot is churlish and noncooperative and will avoid communicating with it.

According to our theory of user involvement, each participant in the communication lies in a different cognitive space before the communication begins. At the commencement of the communication, there should be some means for connecting the different participants' cognitive spaces; this functions as a reference point. Therefore, the participants' cognitive spaces are connected by mutual activities that

are cognitively shared by them. In this paper, we focus on the joint attention as a means to establish user involvement.

Joint attention refers to the interaction where the speaker and the listener cognitively share an object that they focus on. Tomasello suggests that joint attention is divided into three types (Tomasello 1999): check attention (attention to the partner or the object that she/he displays), follow attention (attention to the object that the partner points at or that her/his eye gaze is directed at), and direct attention (making the partner pay attention to the object on which a listener focuses).

Joint attention permits humans to be mutually involved in the cognitive spaces shared by the speaker and the listener. In order to achieve joint attention, the listener has to recognize the speaker's attention and its target by observing her/his behavior, and it should react to this attention appropriately and instantly.

The listener and the speaker employ various types of attention behaviors to express their own attention. We classify attention behavior into four modalities: hand movements, eye gaze, posture, and speech. Hand movements and eye gaze strongly represent attention. Other types of behaviors are also used for representing attention. However, their usage could often be slightly redundant. Attention behaviors are frequently represented with more than one modality or in a repeated fashion. Such redundant behaviors strongly suggest that they are intentional. We assume that there are two types of redundancies in speaker-listener communication: redundancy of modality (multiple modalities of behaviors are used simultaneously) and redundancy of time (repetitive or persistent behaviors are used).

Since redundant behaviors are often intentional, the robot should react rapidly and appropriately to these intentional behaviors of humans.

A human user can adjust herself/himself to a robot with ease if the robot is capable of interpreting redundancies. It is often observed that a human behaves in a redundant manner when she/he cannot communicate with others smoothly. Thus, even if the robot's recognition abilities are insufficient, it can communicate with the human by using a redundant means of communication.

Various types of social skills can be used as a policy for establishing a more advanced communicative reality. Aikawa describes the social skills involved in listening to a partner's speech (Aikawa 2000); these as summarized in Table 1. It is difficult for robots to understand the content of the speech; however, it is possible for them to behave as if they were listening to the speaker through the use of these skills.

Table 1 Social skills for listening to speech

Capable pose	No interruption, no rush
Open question	Prompting, explaining in greater detail
Reflection	Response, repeat, paraphrase, summary
Using nonverbal channels	Posture, gaze, nodding, hand movements
Decoding nonverbal channels	Voice (speed, pitch), emotion, gaze, hand movements

The architecture of the system

We proposed the architecture of a communicative artifact system comprising two layers: the upper layer and the lower layer, as shown in Fig. 1. The lower layer detects events in the environment, and the upper layer coordinates the behavior of the artifact (robot) by handling the evidences provided by the lower layer.

The lower layer functions to establish robust event recognition by focusing on recognizing rhythmic and repetitive gestures and generating immediate responses. Since it is relatively easy to extract periodic behavior even from noisy data and it is important to enable a robot to respond quickly, our approach may result in a robust human–robot interaction.

We use an entrainment mechanism to couple different autonomous communicative artifacts, where entrainment is defined as a type of phenomenon in which the orbits of a human user's gestures and the movements of the robot become synchronized during human–robot interaction. The entrainment-based interaction is described in (Tajima et al. 2004). Through the entrainment-based interaction, a human user can convey her/his tacit intention to a robot by using repetitive gestures.

The upper layer receives the information obtained by the lower layer and determines its behavior. The information necessary for the robot to establish a

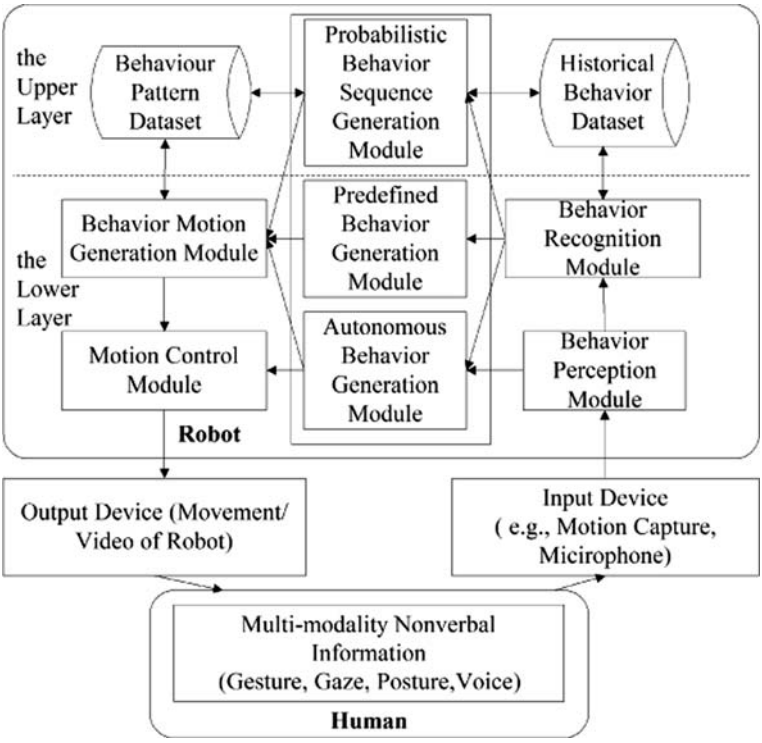

Fig. 1 Architecture of the artifact system

communication process includes body orientation, head orientation, hand position, object position, and pointing direction. In the case of communication between human beings, many social norms are followed. We can extract some patterns from such social norms and call them interaction schema.

An interaction schema describes a typical pattern of interaction between the human and the robot (Hatakeyama 2004). The robot infers the human user's intention and environment situation according to the interaction schema in a manner that it can determine its next operation. Moreover, the flow of communication can be designed by arranging some schemata in a sequence.

Figure 2 presents an example of interaction schema. Sensory data are obtained by measuring the body orientation, gaze direction, and speech as well as the historical record of human–robot interaction. We prefer to use dynamic Bayesian networks (DBN) to implement interaction schema (Tarasenko and Nishida 2006). CM_T denotes the communication mode at time T. B_T represents the behavior of humans at time T. To predict the next communication mode, the system will calculate the probability by a DBN from the human's action, the robot's action, and action records. A DBN comprises the probability distribution function of the sequence of N hidden-state variables $CM = \{CM_0, \ldots, CM_N\}$ and the sequence of observable variables $B = \{B_0, \ldots, B_N\}$, where N is the number of time slices observed. The joint probability is then defined by the following expression:

$$P(\mathrm{CM}, B) = \prod_{t=1}^{N-1} P(\mathrm{CM}_t|\mathrm{CM}_{t-1}) \prod_{t=0}^{N-1} P(B_t|\mathrm{CM}_t) P(\mathrm{CM}_0)$$

The causal relationships between the inputs of humans' behaviors and the outputs of the artifact system are expressed in a DBN. All types of information are integrated by the DBN and used to generate the response behaviors of the system.

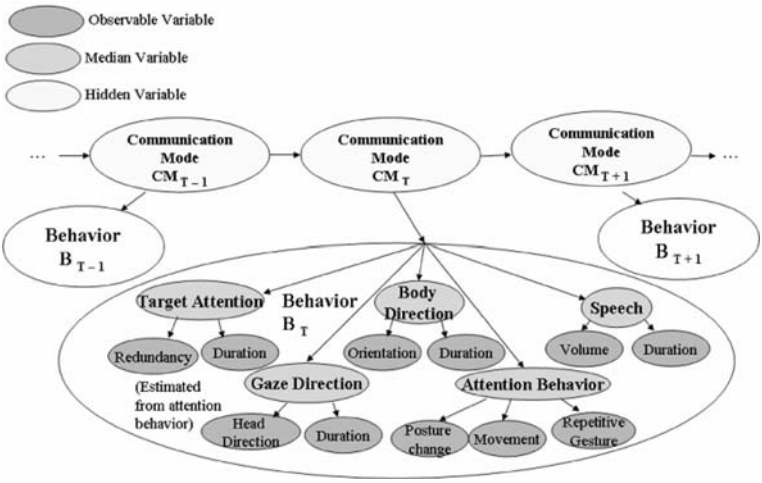

Fig. 2 Interaction schema incorporating DBN

The listener robot

Based on the listener robot system developed by Ogasawara (Ogasawara et al. 2005), we are developing a listener robot system that can participate in a conversation among humans by playing the role of the listener, and record explanatory videos by playing the role of a TV cameraman. The listener robot can listen to a human speaker and record a video as the speaker explains the procedure of assembling and/or disassembling a bicycle to a human listener.

Communication mode

We defined four types of communication modes to describe situations where the speaker, the listener, and the listener robot communicate with each other. As shown in Fig. 3, the four communication modes include: the *talking* mode, the *talking-to* mode, the *explaining* mode, and the *working* mode.

In the *talking* mode, the speaker talks simultaneously with the listener and the robot. In other words, it is not necessary for the speaker to talk to one particular partner. However, in the *talking-to* mode, the speaker talks about something to a specific partner, i.e., the listener or the robot, while watching the listener/robot; in addition, the speaker is involved in the cognitive space that is based on the relations between the communicating partners (the speaker and the listener/robot). On the other hand, when the speaker's gaze is directed primarily toward the target objects to be explained, she/he is in the *explaining* mode. In this mode, the speaker expects the listener/the robot to cognitively share the target. Therefore, the robot should

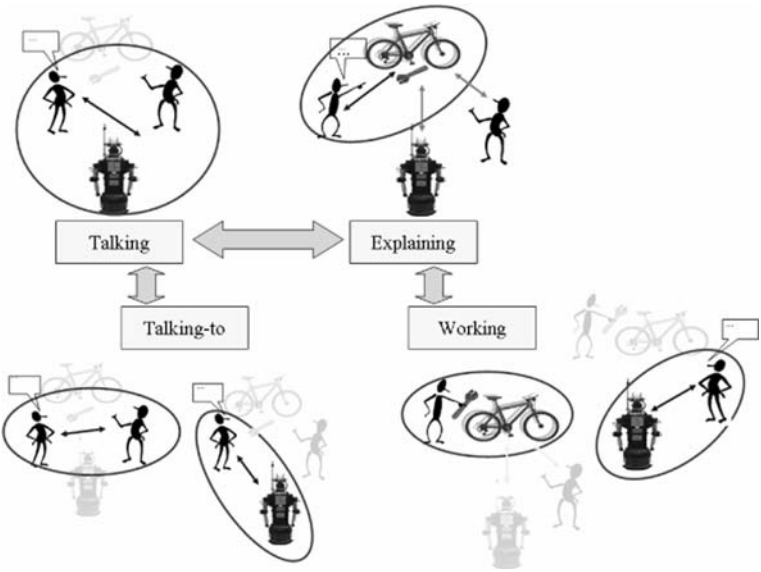

Fig. 3 The communication modes

focus on the target. When the speaker is busy demonstrating the operations to the listener and the robot, she/he may concentrate on her/his work without looking at the listener/robot; we term this mode as the *working* mode. In this case, the speaker appears unconcerned about what the listener/robot is engaged in. Therefore, the robot should remain unobtrusive and not interfere in his/her operation. The four modes are different in the cognitive space that the speaker focuses on. The speaker may intentionally switch between the four modes to establish a communicative reality.

Structure of the listener robot

The structure of the listener robot is shown in Fig. 4. The motion data were obtained from the motion-capture markers attached to the bodies of the humans (including the speaker and the listener) and the robot. The speech data of the humans are input to the robot. The motion and behavior of the robot are generated as the output and feedback to the communication environment.

The following types of information are used in our listener robot system.

Input: Motion-capture data and speech sound.

Information on speaker and listener behavior: The type of speaker and listener behavior, the target object of attention, and the intensity of intention.

Output: Robot behaviors (movement of head, torso, eyes, arms, and approach toward an object) and video data.

The listener robot comprises four modules: attention recognition module, communication mode estimation module, immediate response generation module, and robot behavior decision module. The attention recognition module can

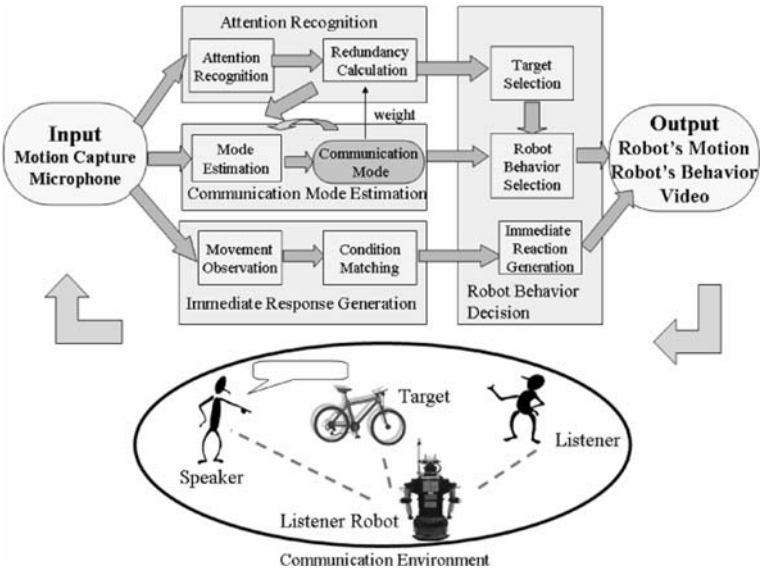

Fig. 4 Structure of a listener robot

recognize the body movements and speech of the speaker or listener; further, it can identify the type of human behavior and the object(s) of attention of the speaker and/or listener. In addition, the communication mode estimation module can estimate the intensity of the intention based on the redundancy of the behaviors. Using these types of information, the robot behavior decision module will select the most confident behavior according to the communication mode. In order to avoid a failure of communication between humans and the robot caused by the slow response of the robot, the immediate response generation module enables quick responses by the robot. As a result, the human will not be confused if the robot is unable to respond within a reasonable time.

Recognition of human behavior

The *attention recognition module* estimates the type of speaker behavior based on the input from the communication environment. In addition, it also estimates the object of the attention when the estimated behavior type is "attention behavior." The motion capture senses the hand movements, the orientation of the torso and face, and the positions of the speaker and the listener. Pointing and repetitive emphasis gestures are recognized by calculating the relative angle/position relations of the shoulder, elbow, and wrist. It is difficult to identify the gaze direction by motion capture. Therefore, gaze direction or body orientation is approximated by the orientation that is determined by two markers fixed at the front and back of the head or body of the human. With regard to speech information, only the volume and duration are recognized to help in determining the communication mode.

Implementation

In early implementation of Ogasawara (Ogasawara et al. 2005), the listener robot can communicate with only one people (a human speaker). However, in current implementation, it can communicate with two people (a human speaker and a human listener). Our implementations used *robovie*, a type of humanoid robot. The robovie can move its eyes, head, as well as its arms, and it can move around the environment by using its wheels. We use a motion-capture device to sense the humans' behaviors. The motion-capture markers are attached to the speaker's and listener's head, arms, and body, and the objects to be pointed at.

Figure 5 shows some examples of communication modes when a speaker explains something to a listener and a listener robot. These pictures illustrate four modes when the speaker explains how to assemble and/or disassemble a bicycle to the listener and the robot.

Figure 5a illustrates an example of the talking mode. In this situation, the robot only needs to continue detecting the users' behaviors and record the scene.

Figure 5b illustrates an example of the talking-to mode. The robot detects the orientation of the user's head and infers whether the user is looking at the listener or itself. If the speaker is looking at the robot, as long as the robot shows any response, the user will naturally feel that the robot is listening to her/him even if it is actually unable to understand natural language.

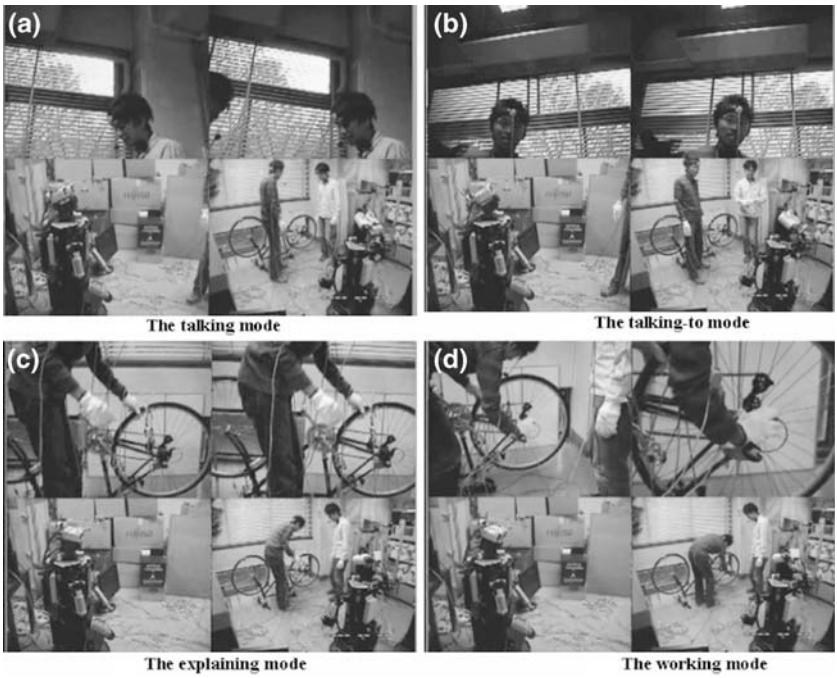

Fig. 5 Examples of the communication modes

Figure 5c illustrates an example of the explaining mode. In this situation, the speaker looks at an object (in this case, the front wheel) and explains how it can be disassembled. The robot will change the photographing angle of its cameras to appear as though it is focusing on the object of the user's attention.

Figure 5d illustrates an example of the working mode. In this case, the speaker concentrates on demonstrating the operation of assembling/disassembling objects, and both the speaker and the listener will remain silent. The robot should remain still and track the users' behaviors. If the speaker focuses on operating certain specific objects, the robot will switch its cameras to the zoom-in mode in order to record better quality video contents.

Since the probability inference which relates to activity log is simple, our primary implementation incorporates activity log as probability variable by using Bayesian networks (BN). We will complete the implementation by replacing BN with more systematic DBN-based model.

Experiment results

In order to obtain a basic assessment of our listener robot system, we conducted a preliminary experiment using a task of assembling and disassembling a bicycle. The users' motions were detected by a motion-capture system (MotionStar wired system of Ascension). Two males in their twenties participated in our experiment. One

Table 2 Recognition rate of communication situations

Communication	Duration time recognized by robot (s)	Duration time recognized by manually (s)	Matching ratio (%)
Conversation (talking and talking-to modes)	170	184	92.39
Explanation (explaining and working modes)	210	279	75.27
Total	380	463	82.07

participant, who had experience in bicycle assembly, acted as the speaker and another inexperienced participant acted as the listener. Both participants wore microphones and had eight motion-capture markers each attached on their bodies (front and back of head, front and back of body, both elbows, and both hands). Four other markers were attached to the front and rear brakes and the front and rear wheels of the bicycle. Video data recorded by two cameras positioned in the environment from front and back angles and that recorded by two built-in eye-cameras of the robot were combined into a composite video data by a four-picture synthesizing device (SG-202II of DAIWA). In the task, the speaker explains how to disassemble and assemble the rear wheel of a bicycle to the listener and the listener robot, first by speech and then by demonstrating the operations. Subsequently, she/he moves to the front wheel and explains and demonstrates the procedure on the front wheel. As a result, we recorded about 8 min of video data and analyzed it using a video analysis tool named Anvil (Kipp 2004). We defined two communication situations in our task: conversation situation (including the talking and talking-to modes) and explanation situation (including the explaining and working modes).

The analysis of the video data was carried out by comparing the duration time recognized automatically by the robot and that recognized manually from the video data. As shown in Table 2, the matching ratio of the conversation situation is 92.39% and that of the explanation situation is 75.27%. A high matching ratio implies that the robot can correctly estimate the communication situation of the humans with a high level of precision. In order to evaluate the effectiveness of the robot when it acts as a TV cameraman, we analyzed the video data to verify the effectiveness of the photography task.

As shown in Table 3, the calculated values of the attention ratio (dividing the attention time of the robot by the attention time of the humans) and the

Table 3 Attention ratio and photographing ratio of attention behaviors and objects

Attention Time of robot (s)	Photographing time of attention objects (s)	Attention time of humans (s)	Attention ratio (%)	Photographing ratio (%)
181	235	289	62.63	81.31

photographing ratio (dividing the photographing time of attention objects by the attention time of the humans) are shown. As the result, the photographing ratio reached 81.31%, and the attention ratio 62.63%. We can infer from these results that the proposed system has a good potential for helping humans communicate knowledge to robots and generate useful knowledge contents in video format. In general, the listener robot system can communicate with humans quite smoothly in the task. We plan to further improve the implementation of the listener robot system by replacing BN with more systematic DBN-based model and conduct additional evaluation experiments to further substantiate the effectiveness of our system.

Conclusion

In this paper, we addressed the problem of establishing a natural communication environment with a robot. We proposed a theory of user involvement by using joint attention and a two-layered human-robot interaction approach. The lower layer improves the robustness of the system by detecting rhythmic and repetitive gestures and by generating immediate responses. The upper layer flexibly controls social interaction by DBN representing social interaction patterns. In order to evaluate our method, we are building a listener robot that can follow and record scenes where a human speaker explains the procedure of assembling and/or disassembling a bicycle to a human listener. Preliminary results are shown to demonstrate how this concept is being implemented on the listener robot.

References

Aikawa M (2000) Techniques of human relation—psychology of social skills, In: Saiensu-Sha (ed)Selection of social psychology, No. 20, (in Japanese)

Hatakeyama M (2004) Human–robot interaction based on interaction schema, Master Thesis, University of Tokyo, Japan (in Japanese)

Kipp M (2004) Gesture generation by imitation: from human behavior to computer character animation, Boca Raton, Dissertation

Matsusaka Y, Fujie S, Kobayashi T (2001) Modeling of conversational strategy for the robot participating in the group conversation, In: Proceedings International Speech Communication Association (ISCA)-EUROSPEECH2001, pp 2173–2176

Okamoto M, Nakano I Y, Nishida T (2004) Toward enhancing user involvement via empathy channel in human-computer interface design, In: Proceedings of international workshop on intelligent media technology for communicative intelligence (IMTfCI 2004), Warsaw, September, pp 129–132

Ogasawara Y, Okamoto M, Nakano IY, Nishida T (2005) Establishing natural communication environment between a human and a listener robot, In: Proceedings of social intelligence and interaction in animals, robots and agent (AISB 2005), conversational informatics for supporting social intelligence and interaction: situational and environmental information enforcing involvement in conversation, England, April, pp 42–51

Reeves B, Nass C (1996) The media equation: how people treat computers, television, and new media like real people and places, CSLI Publications

Tajima T, Xu Y, Nishida T (2004) Entrainment based human-agent interaction, In: Proceedings of 2004 IEEE Conf. on Robotics, Automation and Mechatronics (RAM2004), Singapore, pp 1042–1047

Tarasenko K, Nishida T (2006) Dynamic Bayesian Networks for modelling of alignment and mutual adaptation with communicative robots, In: Proceedins of workshop of social intelligence design (SID2006), pp 39–46

Tomasello M (1999) The cultural origins of human cognition, Harvard University Press, Cambridge
Watanabe T, Ogawa H (2001) InterRobot for human interaction and communication support. In: Proceedings of world multi-conference on systems, cybernetics and informatics (SCI2001) pp 466–471