Establishing Natural Communication Environment between a Human and a Listener Robot

Yoshiyasu OGASAWARA[*]
Masashi OKAMOTO[*]
[*]Graduate School of Information Science and Technology, the University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
yoshiyas@kc.t.u-tokyo.ac.jp, okamoto@kc.t.u-tokyo.ac.jp
Yukiko I. NAKANO[†]
[†]Research Institute of Science and Technology for Society, Japan Science and Technology Agency
Atago Green Hills MORI Tower 18F, 2-5-1 Atago, Minato-ku, Tokyo, 105-6218, Japan
nakano@kc.t.u-tokyo.ac.jp
Toyoaki NISHIDA[‡]
[‡]Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
nishida@i.kyoto-u.ac.jp

Abstract

The progress of technology makes familiar artifacts more complicated than before. Therefore, establishing natural communication with artifacts becomes necessary in order to use such complicated artifacts effectively. We believe that it is effective to apply our natural communication manner between a listener and a speaker to human-robot communication. The purpose of this paper is to propose the method of establishing communication environment between a human and a listener robot. In our method, their common intention is formed by joint attention and redundancy of behaviour.

1 Introduction

In recent years, the functions of familiar artifacts, such as an electric appliance and a personal computer, gets complicated as they are being advanced rapidly. To fully use the functions of such artifacts without giving a burden to the user is difficult with the present method, which the user has to learn in advance. The user should be able to tell the artifacts what she wants to do, that is, her ‘intention’.

But, the user's intention does not always appear from the start. In most cases it is gradually appearing and becoming clear during the process of the communication. Therefore, it is important for the user to establish natural communication with the complicated artifact.

However, there are many examples where the natural communication is not established between humans and artifacts. One of the typical examples is making a video letter with a video cam. In this case the speakers in video letters often present unnatural way of speaking and behaviour. But, they will not do such a way of speaking when they talk to their close persons. We assume that this difference is caused by whether natural communication between a speaker and a listener is established or not.

The human listener responds to the speaker's behaviour with various ways. The listener implicitly conveys his listening attitude to the speaker through his responses or gestures. These behaviours and responses establish the communication and let the speaker feel relaxed. On the contrary, in making a video letter, the video camera is placed at the listener's position instead of a human listener. Then the communication is not established because the camera does not response to the speaker at all. Therefore, the speaker feels stressed and cannot behave as usual.

The purpose of this research is to build a listener robot whose natural behaviours as a listener allows its user to speak and behave in unrestrained ways. And then we propose the method of establishing the environment of natural communication where a human explains to the robot with gestures such as pointing. As a result it will be shown how the users get to exercise the functions of complicated artifacts effortlessly.

2 Related works

Many researches have been made on building a natural communication robot. Some of them discuss the matter in view of the listener's behaviour.

Watanabe and Ogawa (2001) developed InterRobot for the purpose of smoothing the voice conversation between the remote places. When the human speaks to InterRobot, it generates the gestures such as nodding from the user's speech, and it reacts as if it were listening to the user. On the other hand, speech information is sent to the remote place, and InterRobot duplicates the speaker's gestures conjectured from the information.

Kismet, which Breazeal and Scassellati (1999) developed, is one of the robots aiming at the human-like behaviour. Kismet can recognize a thing in its sight to which it should pay attention by vision processing technology. It can turn its neck and eyes toward a human face or an object that moves quickly.

Although these works try to make robots behave like a human, what he is talking about is not taken into consideration. Therefore, it is difficult for the human to explain about some topic to the robot in a natural way.

Ozeki et al. (2001) developed the video contents creation supporting system. In this system, the user speaks toward a set of cameras with specific gestures and words for operating cameras.

In this method, it is difficult to do the usual way of speaking because there is no communication with the artifact. If the system is able to communicate with users, it is expected that users can speak naturally. Such ability will make it possible for a human to use the system with ease, even if the system is updated and its function becomes complicated.

3 Natural communication with artifacts

In our research, we aim to establish natural communication between a human and a listener robot. In this section, we introduce the idea of 'User involvement' (Okamoto et al., 2004) as the basic framework in designing the computer-mediated communication environment. In enhancing the user involvement in human-to-robot communication, joint attention is significant for the communicative reality to be established there. Moreover, social skills for communication are also introduced to smooth the communication between a human and a listener robot.

3.1 User involvement

There are many discussions about realizing natural human-computer interaction. In our research we focus on the idea of 'User involvement' that Okamoto et al. (2004) put forward. User involvement means the cognitive way humans willingly engage in the interaction with computers, or the way in which humans are, on the contrary, forced to be involved in a virtual world which computers display or in a human-to-robot communication. The requirements are considered as follows:

- **Cognitive/Communicative reality should be achieved:** The user should feel the virtual object/world, or the human-to-computer interaction as "real".
- **Two (or more) cognitive spaces should be linked:** The user should move in and out smoothly at least two cognitive spaces such as his/her viewpoint (here) and what he/she sees (there).

Our goal is to achieve the communicative reality in the main so as to enhance the user involvement by implementing natural responses and reactions as a good listener into the listener robot. For it is difficult to fully establish the cognitive reality using a robot, in that current humanoid robots do not have so sufficient appearances or facial expressions as to make humans feel as if they the robots were living. Moreover, as Reeves and Nass (1996) points out, humans are likely to behave toward artifacts as if they were humans. Therefore, if a robot reacts to humans in unnatural ways, then they will assume it is churlish and non-cooperative and will not keep communicating with it.

From the point of the view of the user involvement, it can be said that each of the participants in communication originally lies in a different cognitive space before the communication begins. But, once the communication starts, there has to be something to connect those cognitive spaces, which functions as a reference point for one participant to access another. Then the communicative reality for the participant is achieved.

In human communication, what connects the participants' cognitive spaces is that which are cognitively shared in the participants, such as exchanging verbal information, establishing eye contact, matching action and reactions, joint attention, and so on. We believe that those factors that enable human communication will be applied to human-robot communication as well.

We focus on the 'joint attention' in particular in order to establish natural human-robot communication environment. Specifically, the speaker-listener communication using a listener robot is described.

3.2 Model of speaker-listener communication environment

Figure 1 illustrates the model of the speaker-listener communication environment. A speaker and a listener keep exhibiting their behaviours, such as gestures or utterances, to each other in the communication. During the process, one participant's intention is approaching that of the other, and then the common intention is formed. The common intention here means rather the 'intentionality', which Dennett (1987) suggests, than the 'intention' as used in a usual context. For the common intention concerns what is mutually aimed both by the speaker and the listener.

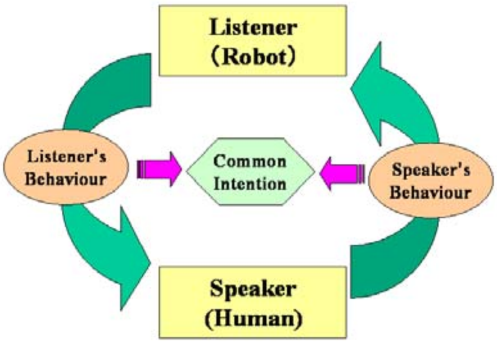


Figure 1: Model of speaker-listener communication environment

Take the explanation task for communication example. The state of the communication environment changes depending on what is being explained, or what is being attended to. When applying the model to the human-robot communication with a listener robot, the robot needs to change their behaviours according to the state of the communication environment. Therefore, in order to decide the robot's behaviour at each moment of the communication process, it is necessary to fully analyze what comprises the given domain of explanation.

3.3 Joint attention

Joint attention is the interaction where a speaker and a listener cognitively share an object that they attend to. For example, the listener looks at the thing the speaker points at. In human communication, these interactions are being done unconsciously and make the participants feel the communication natural. In fact the function of joint attention was implemented into such a humanoid robot as Kismet (Breazeal et al., 1999) and Infanoid (Kojima, 2000), and was proved to be effective for establishing natural human-robot communication. Therefore, joint attention is also one of the important requirements to establish communicative reality between the speaker and the listener.

In view of the user involvement, the joint attention leads humans to be mutually involved in the same cognitive space in the communication, which is the overlapped part of both cognitive spaces of speaker and listener. As a result, the joint attention enables the listener to accesses the speaker's attention object through the joint attention as a reference point, and vice versa. Therefore, joint attention achieves the smooth transition among the cognitive spaces of the participants, which helps to establish the communicative reality for them.

3.4 Social Skills

We propose the idea to use the social skills, as a policy for establishing more advanced communicative reality. Social skill is the theory to establish human relations effectively in view of the communication techniques. Aikawa (2000) describes the social skills for listening to the partner's speech. Table 1 summarizes them.

Table 1: Skills for listening to speech

Capable Pose	No interruption, No rush
Open Question	Prompting, Explaining more in details
Reflection	Verbal response, Repeat, Paraphrase, Summary
Using Non-verbal Channel	Posture, Gaze, Nodding, Distance, Hand's movement
Decoding Speaker's Non-verbal Channel	Voice (pause, speed, pitch), Emotion, Gaze, Hand's movement

In these skills, verbal response and nodding are important for building the listener robot. Many people have experienced that the listener's nodding improves the rhythm of the speaker's speech. It is difficult for robots to understand the contents of the speech, but it is possible for them to behave as if they were listening to the speaker through using these skills.

4 Analysis of speaker-listener communication

In this section we analyze the characteristics of human communication between a speaker and a listener in order to define the appropriate behaviours of our listener robot. First we describe the human behaviours for establishing joint attention according to psychological studies. Secondly we analyze the videos which captured explanatory scenes of humans and show the observations of the videos.

4.1 Behaviours for establishing joint attention

As shown in the previous section, joint attention is one of the significant factors that constitute speaker-listener communication.

Tomasello (1999) suggests that joint attention is divided into the following 3 types according to the development levels of infants:

- **Check Attention:** Attention to the partner, or the object he shows.
- **Follow Attention:** Attention to the object that the partner points at or his eye gaze turns to.
- **Direct Attention:** Making the partner pay attention to the object with voice, eye gaze, and so on.

This classification indicates that it is necessary for establishing joint attention to properly react to the behaviours such as showing by hand, pointing by finger and turning eye gaze.

Moreover, Clark (2003) explains about the functions of such behaviours as *pointing* and *placing* in human communication. Among them the behaviours for 'directing-to' an object are classified into the following categories:

- *Pointing* (finger)
- *Sweeping* (arm)
- *Tapping* (finger, foot)
- *Nodding* (head)
- *Turning* (torso, face)
- *Eye Gazing*
- *Speaking* (frequently accompanied by head and face)

In particular *eye gazing* is considered to be the most attention-getting behaviour. Additionally, where an object is placed affects the listener's attention. For instance, placing an object onto a desk or in front of the hearer gets his attention much.

Therefore, each of these behaviours should be counted as attention behaviour of a speaker.

4.2 Observations of explanatory movie

We observed an explanatory video movie so as to confirm that those attention behaviours are actually used for explanatory communication. The data we used is 40 minutes of an educational material video for DIY (Do It Yourself) in which a professional instructor explains how to use machine tools against the TV camera.

As a result of the observation, it was proved that most of the attention bahaviours were actually used in explaining scenes. In particular, pointing, showing and eye gaze were most frequently used, when the camera also focused on the object.

Moreover, some of the attention behaviours were frequently used simultaneously (e.g. Showing + Pointing + Gazing + Speaking "*This is*..."). During the large part of the instruction the instructor was attending either to an object to be explained or to the camera in front of her. We also found that the direction of her gaze continuously changed between the object and the camera at short intervals while she was speaking.

4.3 Analysis of listener's response behaviours against speaker

Since the explanatory task observed in the previous section is toward a TV camera alone, it is different from the actual explaining communication between a human speaker and a listener. We thus made two movies of explanatory scenes between humans and analyzed the listener's behaviours against the speaker's ones using a video annotation tool, Anvil[1] (Figure 2).

Figure 2: The video data of explanatory scene

The task to be explained by the speaker was how to assemble a piece of furniture (a metal rack). The speaker was one of the authors and the listeners were two students.

Table 2: Analysis of listener's behaviours

	Listener's response to speaker's gaze		
	Joint attention	Gaze toward speaker	Nodding toward speaker
Listener 1	76.4	85.3	64.7
Listener 2	84.7	71.4	47.6

(%)

Table 2 shows the result of the analysis. When the speaker attends to an object, the listener attends to it at more than 75%. Therefore, joint attention between the speaker and the listener is achieved at high frequency. Moreover, when the speaker turns his gaze on the listener, the listener turns his gaze

[1] http://www.dfki.de/~kipp/anvil/

back to the speaker at more than 70%. In many cases the listener gives a nod in concurrence with his gaze, but its frequency differs greatly in individuals.

Among those exchanges occurred in the communication the most frequent behaviour transition during a short period is as follows:

1. The speaker turns his gaze on the listener.
2. The listener turns his gaze back to the speaker.
3. The listener gives a nod (or does nothing).
4. The speaker looks at the object to be explained.
5. The listener looks at it.

We assume that this transition occurs because the speaker wants to confirm the listener's attention.

In addition, the following characteristics were commonly observed regarding the two listeners:

- Among the speaker's behaviours, showing the object by hand and turning gaze especially attract the listener's attention.
- The listener usually attends to the object after the speaker's multiple behaviours (e.g. showing + gaze).
- When the speaker moves or changes his posture, the listener attends to the speaker himself instead of the object.

4.5 Summary

The observations suggest that the speaker-listener communication in explanatory task has the following characteristics:

(1) On attention behaviours:
- According to the speaker's attention with pointing, gaze, or posture, and then **the listener attends to the same object**.
- Among the attention behaviours **showing by hand**, **pointing** and **gaze** are the most affective ones, and are **often used simultaneously**.

(2) On communication modes:
- There are different communication modes in explanation: (a) the speaker **attends to an object** and explains about it to the listener, (b) the speaker **attends to the listener** and talks to him, (c) the speaker glances at the listener to **confirm the listener's response**.
- In other words, **the communication mode changes according to each relation among the speaker, the listener and the object**.

5 Requirements for a listener robot

In this section we describe the requirements for building a listener robot based on the observations and the analysis in previous sections.

5.1 Establishing joint attention

As shown in Section 4, it is frequently observed that both the speaker and the listener attend to the same object in explanation task. Therefore, implementing the ability of joint attention into a listener robot is required.

The correspondence of the listener's proper reaction with the precedent actions of the speaker is necessary for the communication between a speaker and a listener to be established. It is through this repetitive process that natural speaker-listener communication is achieved. Among the correspondent action and reaction couplings, the most fundamental and effective one is seemingly joint attention.

In order to achieve joint attention, the listener has to recognize the speaker's attention and its target from the observation of the speaker's behaviours, and should react to the attention appropriately and instantly.

Moreover, it is also essential for establishing joint attention whether the attention behaviours are put out by the speaker intentionally or not. For reacting against non-intentional behaviours would become unnatural in communication. Therefore, we proposed the method of detecting the intensity of the speaker's intention based on the redundancy of the behaviour.

5.2 Modality of attention behaviour

Table 3 summarizes attention behaviours in view of each modality based on the observations in Section 4.

Table 3: Modality of attention behaviour

Modality	Behaviours
Hand movements	Pointing, Grasping, Showing
Eye gaze	Direction (head, eyes)
Posture	Approaching, Direction (body) Standing-up, Sitting-down
Speech	Deictic words, Verbal response, Mentioning

A listener and a speaker use these types of behaviours to express their own attention or to interpret the other's. In Table 3, hand movements and eye gaze are assumed to strongly represent attention. If a listener robot is able to recognize such attention behaviours of the speaker, then the establishment of joint attention will be easier.

The other types of behaviours are also used for representing attention. However, they are often used with a certain degree of redundancy. The redundancy of attention behaviour is described in the next subsection.

5.3 Redundancy of attention behaviour

We observed that attention behaviours are frequently represented with more than one modality or in repeating fashion. Such redundant manners strongly suggest that those behaviours are intentional. In other words, the redundancy of behaviour manner helps a listener to recognize the speaker's intention.

We propose that the redundancy of attention behaviours should be applied to the design of natural communication environment using a listener robot. The speaker's redundant behaviours strongly suggest the intentionality of his behaviours. Recognizing the redundancy enables the listener robot to easily understand the speaker's intention. The effectiveness of applying the informative redundancies to an agent's sensors and actuators is also suggested by Pfeifer and Scheier (1999).

It is assumed that there are the following two types of redundancy in the speaker-listener communication:

- **Redundancy of modality:** Using multiple modalities of behaviours simultaneously
- **Redundancy of time:** Using the repetitive or persistent behaviours

Since the redundant behaviours are often intentional, communicative robots should react rapidly and appropriately against those intentional behaviours of humans. Conversely, the robots also can convey strong intention to humans using redundant actions or behaviours. This effect of redundancy is discussed in Tajima (2004) through the human-robot communication experiment.

A human can adjust himself to a robot with ease if the robot can interpret redundancy. It is often observed that a human behaves in more redundant ways when he cannot communicate with others smoothly. Thus, even the robots with insufficient recognition abilities can communicate with a human as long as he is willing to use the redundant ways of communication.

5.4 Modes of Communication

Intentional attention behaviours are not always used in the process of explanatory communication. It is required for a listener robot to behave properly even if there are no such intentional behaviours of a speaker.

The analysis in Section 4 suggests that the proper behaviours of a listener robot should be determined depending on the relations between the speaker, the listener and the target. We call that a communication mode. According to each of the communication modes the listener robot should be able to change its behaviours properly. The modes of communication are classified into the following four types in view of the speaker's attention (also see Figure 3):

- Talking-*to* mode:
- Talking-*about* mode
- Confirming mode
- Busy mode

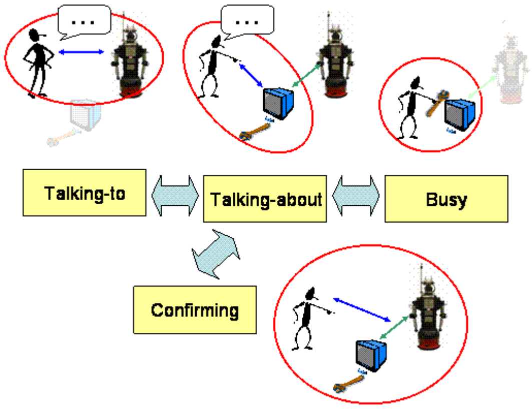


Figure 3: The modes of communication

In general, the speaker engaged in the explanation task attends to either the listener or the target to be explained. In the '**talking-*to***' mode, the speaker is mainly watching the listener and is involved in the cognitive space based on the relation between the speaker and the listener. As the speaker in the 'talking-*to*' mode expects the listener to be involved in the same conversation, the listener should pay attention to the speaker himself.

On the other hand, when the speaker is mainly watching the target to be explained, he is in the '**talking-*about***' mode. In this mode, the speaker expects the listener to cognitively share the target. Therefore, the listener should attend to the target in turn.

Additionally, in the cases where the speaker switches his gaze between toward the listener and toward the target, he is interested in the relation of the listener and the target, that is, whether or not the listener is paying attention to the target. We thus call the mode as the '**confirming**' mode in that he tries

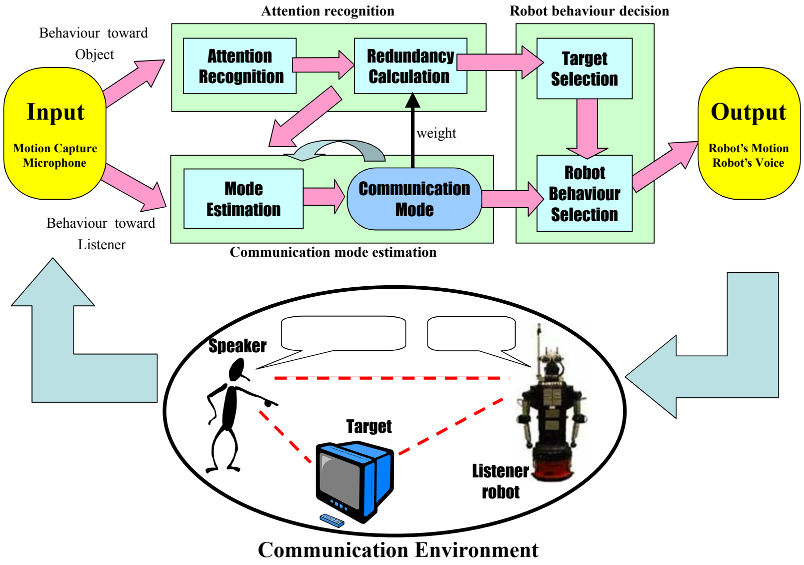


Figure 4: Architecture of listener robot

to confirm the listener's attention. In the confirming mode, the listener should turn his gaze back to the speaker and respond with nods or verbal responses.

The '**busy**' mode is the mode when the speaker is devoting himself to his work without talking to his listener. In this mode the speaker is attending more to the target than in the talking-*about* mode, and tends to ignore the listener. This situation is not favorable for explanation task, but frequently occurs when the speaker is not skillful. A listener robot can keep attending to the target during the busy mode, which is one of the advantages for using robots.

These modes of the speaker's attention differ in which cognitive space the speaker is involved. The speaker creatively uses the multiple modes to establish communicative reality. If the listener robot does not appropriately react according to the specific communication mode, the speaker will not be able to smoothly switch these modes, and then he will not feel the communication with the robot as real.

6 Constructing a Listener Robot

In this section, we describe the implementing method of the attributes of establishing natural speaker-listener communication environment, suggested in Section 5, into the listener robot.

6.1 Target

In this research, we construct a listener robot, which is designed to participate in such human-robot communication environment as the situation where a human explains the procedure of assembling a piece of furniture or an appliance toward the listener robot.

This task of explaining the procedures of assembling furniture implies the basic behaviours that are usually used by the speaker and the listener. Therefore, constructing a listener robot involved in the task will be also helpful to solve the problems of many other explanation tasks.

6.2 Hardware

In constructing a listener robot, we use Robovie[2], a humanoid robot. Robovie is nearly as tall as a human and can move its hands, head, and eyes. Additionally, it can make a move with its wheels.

The motions of the speaker's body or the tools he uses are recognized via the motion capture. The markers of the motion capture are attached to the speaker's head, arms, body, and the objects to use or to point at in his explanation. In addition, one microphone is used to measure the speech sound volume of the speaker.

6.3 Architecture of listener robot

The architecture of the listener robot is shown in Figure 4.

The robot receives motion capture data of the speaker's behaviours and the speech sound as input. As the result of processing the data, the robot outputs bodily expressions as feedback to the communication environment.

[2] http://www.mic.atr.co.jp/˜michita/everyday-e/

- **Input:** Motion capture data, and speech sound.
- **Output:** Robot behaviours (head, eyes, arms, head nod, voice, and body direction)

The following subsections describe the methods used in each module in the robot system architecture.

6.3.1 Attention recognition

In order to recognize the attention behaviours of a speaker, the listener robot processes each data from respective markers attached to all the possible targets. The details of the process are described below.

First, the confidence for attention toward each target is calculated from the positions of the motion capture markers to decide the following behaviours:

- Eye gaze (head direction)
- Pointing
- Grasping
- Repetitive hand gestures (e.g. tapping)
- Physical relationship with objects (distance, body direction)

Respective bahaviours are recognized based on simple calculation of such as distance and angles between markers. For example, gaze direction is estimated from the directions of two markers attached on the speaker's head. Grasping is recognized by calculating the distance between hand and object. The precise direction of eyes is difficult to recognize from motion capture information. Therefore, we are not concerned it in this listener robot.

Repetitive gesture of hands is calculated by employing methods proposed by Tajima (2003) and Anuchitkittikul (2004). Their method detects human's repetitive gestures by using an autocorrelation function and the Fast Fourier Transform (FFT).

The calculation of the confidence for respective gestures is based on the relational network of sensor data and behaviours as illustrated in Figure 5. Outputs from this process are calculated by using the method proposed by Hatakeyama (2004). This theory applies Bayesian Network to the processing of sensor information and the decision of robot's behaviour. In this method, inputs and outputs are represented as random variables. The causal relationships between inputs and outputs are expressed as a conditional probability table.

This method has an advantage over other methods that describe the relationship between input and output as rules, because it is more robust against the noisy input and the changes of communication environment.

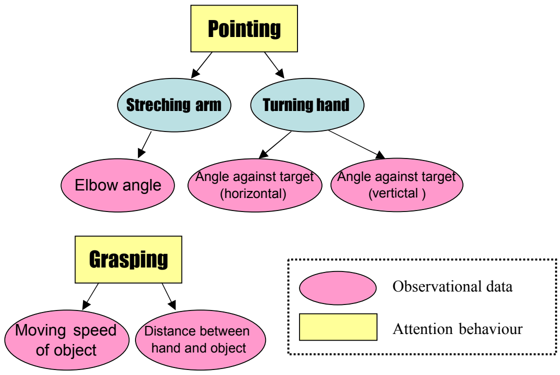


Figure 5: Network for attention recognition (partial)

Each node of the network was decided by us based on the observations in Section 4.

Furthermore, the confidence values of the respective behaviours are regarded as scores, and then the scores are weighed and added up into the redundancy of attention behaviours toward the target (Redundancy of modality).

Weighing scores depend on the current communication mode and the duration of behaviours (Redundancy of time). For example, in the Talking-*to* mode, weighed values are less than in any other mode because gazing attention to the target does not frequently occur in the Talking-*to* mode.

6.3.2 Communication mode estimation

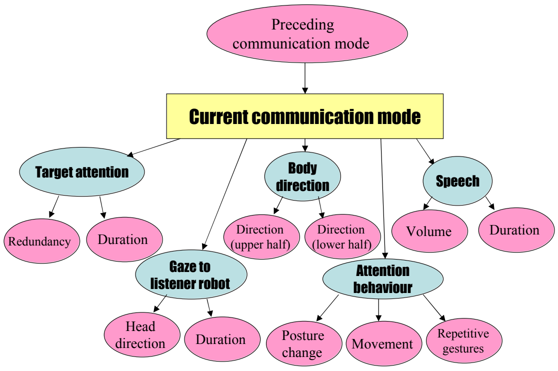


Figure 6: Estimation of communication mode

The communication modes suggested in Section 5 are classified according to what relation is established between the speaker, the listener and the target. Our listener robot decides each communication mode based on the speaker's behaviours.

Which communication mode the speaker and the listener are involved in is estimated from the following inputs using the similar algorithm to that of attention recognition (also see Figure 6):

- Target attention (with the largest redundancies)
- Gaze to the listener robot
- Body direction toward the listener robot
- Attention behaviour (e.g. posture change)
- Speech (volume, duration)
- Preceding communication mode

The general conditions for recognizing each communication mode are as follows:

- **Talking-*to***: with frequent speech, speaker's body and face toward the listener
- **Talking-*about***: with attention behaviours and gaze to the target
- **Confirming:** with momentary gaze toward the listener during target attention
- **Busy:** with no speech, with continuous attention to the target

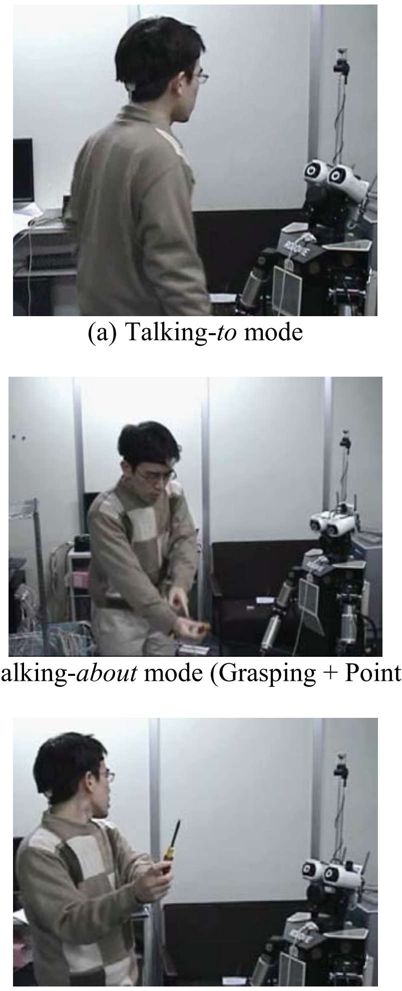
6.3.3 Robot behaviour decision

The listener robot's behaviours are decided by if-then rules based on the redundancy of attention behaviour and the communication mode. The object with the largest redundancy is selected as the target for the listener robot's attention.

The manners of the listener robot's behaviours are summarized as follows:

- **Talking-*to***: turns the head to the speaker and randomly nods when the speech is aborted.
  - ➢ **Low redundancy:** *no attention*
  - ➢ **Medium redundancy:** *gaze attention in a short period*
  - ➢ **High redundancy:** *slight head move and gaze attention in a short period*
- **Talking-*about***: uses various methods.
  - ➢ **Very low redundancy:** *no attention*
  - ➢ **Low redundancy:** *gaze attention*
  - ➢ **Medium redundancy:** *head attention after gaze attention*
  - ➢ **High redundancy:** *prompt head attention*
  - ➢ **Very high redundancy:** *head move and verbal responses*
- **Confirming:** takes a glance at the speaker, nods with verbal responses, and then returns to the original state. Nodding and verbal responses are randomly produced.
- **Busy:** in the same fashion as talking-*about* mode, but with no utterances,

6.4 Examples of explanation for the listener robot


(a) Talking-*to* mode

(b) Talking-*about* mode (Grasping + Pointing)

(c) Confirming mode

Figure 7 : Example scenes of explanation for the listener robot

Figure 7 shows examples of explanation for the listener robot. These pictures illustrate three speaker's modes when the speaker explains how to assemble the metal rack using tools.

Picture (a) shows that the speaker is talking to the robot and the robot turns its head to the speaker. In picture (b) the speaker attends to a tool with multiple behaviours. Then the robot turns its head to the tool, when the joint attention between the speaker and the robot is established. In picture (c) the speaker looks at the robot so as to confirm its reactions during his attention to the tool. At the time the robot glances at the speaker with its eyes and nods to the speaker.

As a result, the listener robot efficiently recognized speaker's intentions using the redundancy of speaker's natural behaviours. Moreover, considering communication modes helped a lot to realize the proper reactions of the robot for explanatory task.

7 Conclusion

In this paper, we proposed the method to establish the natural communication environment with the artifact such as a robot. We developed the listener robot applying this method in the explanation of assembling furniture. The listener robot established natural joint attention with a human speaker using the redundancies of attention behaviour.

In the future, we will examine psychological burden of the user in the explanation task with the listener robot. Additionally, we will analyze more detailed behaviours of human speakers and listeners and define more proper communication modes and robot's behaviours in each mode.

Finally, we will develop the system supporting creation of the video contents using this listener robot. If this robot's behaviour makes a speaker relaxed, the speaker is expected to be able to create good video contents.

References

Mitsuru Aikawa. *Techniques of Human Relation – Psychology of Social Skills (in Japanese)*. Number 20 in Selection of Social Psychology. Saiensu-Sha, 2000.

Burin Anuchitkittikul, Masashi Okamoto, Hidekazu Kubota, and Toyoaki Nishida. Gestural interface for the creation of personalized video-based content. In *Proceedings of the 2nd International Conference on Information Technology for Application (ICITA 2004)*, China, 2004.

Cynthia Breazeal and Brian Scassellati. A context dependent attention system for a social robot. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99)*, 1146–1151, Stockholm, Sweden, 1999.

Herbert H. Clark. Pointing and placing. In *Pointing: Where Language, Culture, and Cognition Meet*. Lawrence Erlbaum Assoc Inc., 2003.

Daniel C. Dennett. *The Intentional Stance*. The MIT Press, 1987.

Makoto Hatakeyama. Human-Robot Interaction based on Interaction Schema (*in Japanese*). Master's thesis, Graduate School of Information Science and Technology, The University of Tokyo, 2004.

Hideki Kozima. Infanoid: An experimental tool for developmental psycho-robotics. In *Proceedings of International Workshop on Developmental Study*, Tokyo, 2000.

Yoshiyasu Ogasawara, Takashi Tajima, Makoto Hatakeyama, and Toyoaki Nishida. Human-robot communication of tacit information based on entrainment (in Japanese). In *Proceedings of The 18th Annual Conference of the Japanese Society for Artificial Intelligence*, 2004.

Masashi Okamoto, Yukiko I. Nakano, and Toyoaki Nishida. Toward enhancing user involvement via empathy channel in human-computer interface design. In *Proceedings of IMTCI*, 2004.

M. Ozeki, Y. Nakamura, and Y. Ohta. Camerawork for intelligent video production –capturing desktop manipulations. In *Proceedings of International Conference on Multimedia and Expo*, 41–44, 2001

Rolf Pfeifer and Christian Scheier. *Understanding Intelligence*, Bradford Books, 1999.

Byron Reeves and Clifford Nass. *The Media Equation: How People treat computers, television, and new media like real people and places*. CSLI Publications, 1996.

Takashi Tajima and Toyoaki Nishida. Manual-less interaction based on synchronization and modulation (*in Japanese*). In *Proceedings of IEICE Human Information Processing*. The Institute of Electronics, Information and Communication Engineers, December 2003.

Michael Tomasello. *The Cultural Origins of Human Cognition*. Harvard University Press, 1999.

Tomio Watanabe and Hiroki Ogawa. InterRobot for human interaction and communication support. In *Proceedings of world Multi-conference on Systems, Cybernetics and Informatics (SCI2001)*, 466–471, 2001.