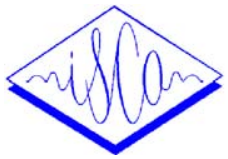

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

Multi-person Conversation via Multi-modal Interface — A Robot who Communicate with Multi-user —

Yosuke Matsusaka, Tsuyoshi Tojo, Sentaro Kubota, Kenji Furukawa,
Daisuke Tamiya, Keisuke Hayata, Yuichiro Nakano and Tetsunori Kobayashi
School of Science and Engineering, Waseda University
3-4-1, Okubo, Shinjuku-ku, Tokyo 169, JAPAN
yosuke@tk.elec.waseda.ac.jp

10.21437/Eurospeech.1999-387

ABSTRACT

This paper describes a robot who converses with multi-person using his multi-modal interface.

The multi-person conversation includes many new problems, which are not cared in the conventional one-to-one conversation: such as information flow problems (recognizing who is speaking and to whom he is speaking / appealing to whom the system is speaking), space information sharing problem and turn holder estimation problem (estimating who is the next speaker).

We solved these problems by utilizing multi-modal interface: face direction recognition, gesture recognition, sound direction recognition, speech recognition and gestural expression. The systematic combination of these functions realized human friendly multi-person conversation system.

1 INTRODUCTION

Multi-person conversation means a conversational situation in which many person talk to each other sharing a common topic. All of usual conversation systems deal with only one-to-one conversation, in which the system talks with only one user. However, if we desire more natural information terminals in our life-space, the multi-person conversation ability becomes indispensable. Now we are developing a robot named ROBITA who acts as an information servant for us (see Fig. 1). In this situation, it is desired for the robot to join our natural multi-person conversation to be a natural partner of us.

The multi-person conversation system has to treat many problems, which are not cared in the conventional one-to-one conversation. In order to follow the conversation correctly, the robot has to recognize information flow (that means who is speaking and to whom he is speaking). And, in case of the robot himself is speaking, he has to appeal his intended information flow (to whom he is speaking). The robot also has to deal with the interruption during the talk. Furthermore, the robot has to recognize the person to whom all conversation participants focus their attention, because such a person likely to speak next.

To solve these problems, we gave the robot multi-modal information recognition abilities and non-verbal body expression abilities. Multi-modal information recognition abilities include gesture and face direction recognition via image data, sound source direction recognition via sound data, and conversational speech recognition via speech data. Non-verbal body expressions include eye contact with face and eye direction control, and gesture expression with hands.

In this paper, the detail of the each sub-functional system and the framework of the total system are described.

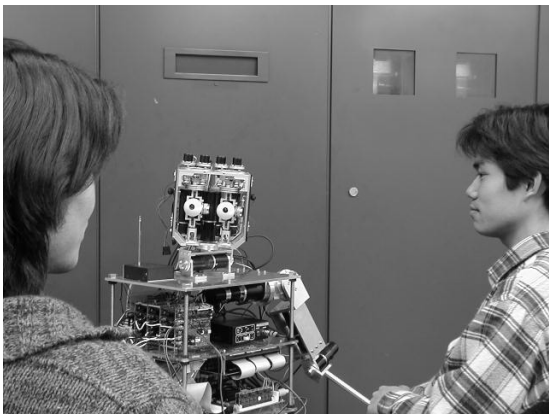

Figure 1: Conversation robot ROBITA

2 PROBLEMS IN MULTI-PERSON CONVERSATION

In this section, we point out the new problems in the multi-person conversation.

2.1 Information flow problems

One big problem of multi-person conversation is the information flow problem: the system has to recognize who is speaking and to whom he is speaking. In the usual one-to-one conversation system, there are no ambiguities on the information flow: input sound to the system is always from the only one person who fronted on the system, and the message on the sound is always to the system. The expected system's behavior is very simple: generate simple answer for the received message and send it to the person fronted on the system.

In the multi-person conversation robot, however, any person can be the speaker. In order to recognize the conversational situation correctly, the robot has to recognize who uttered the sound. The expected behavior of the robot may be different according to the speaker of the message. To realize this function, we prepared sound source direction recognition (described in section 3.2), and face recognition (described in section 3.3). The robot uses sound source recognition to detect which person utters the sound. The robot uses face and speaker recognition to detect who is the speaker.

The robot also has to recognize the intended receiver of the message: to whom the speaker intended to send the message. If the message is to the robot, he has to reply. While, if the message is to the other person, the robot may have to ignore the message. To realize this function, we prepared face direction recognition (described in section 3.3). The robot decided that the direction of the speaker indicates his intended receiver of the message.

While, when the robot utter some messages, another information flow problem arises. The robot has to appeal to whom he intended to send his message. If the intended receiver is ambiguous, the participants of the conversation are bewildered. Such situation never realizes natural rhythmical conversation. To realize this function, we prepared face direction control (described in section 3.3) and eye contact. Eyes are very important to show his attention as well as to see something. The robot directs his face to the intended receiver of the message. By this action of the robot, the all participants can recognize to whom the robot intended to send the message.

2.2 Space information sharing

In the multi-person conversation, all participants share the common space. Therefore, some topics in the conversation include space information. In that case, it is natural for them to use demonstrative pronoun and demonstrative gesture. It is rare case to use direct words to represent places like "Near the bookshelf" . People often point to the place by their fingers and just say "Over there.". To realize the expression of the robot involving demonstrative pronoun and gesture, we prepared gesture expression and space information management ability (described in section 3.6). To recognize the expression of the users involving demonstrative pronoun and gesture, we prepared gesture recognition (described in section 3.4).

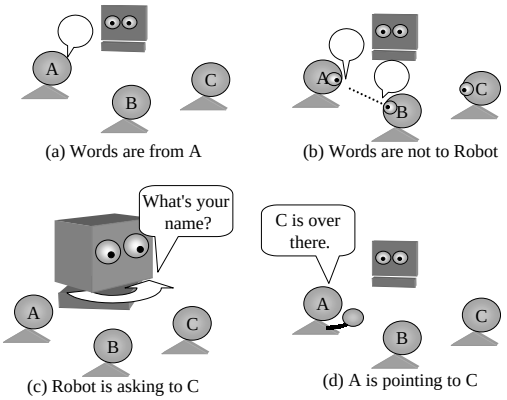

Figure 2: Multi-modal information in multi-person conversation

2.3 Turn holder/focusee control and estimation

If all the participants of the multi-person conversation speak disorderly, the conversation will not go on. So we make a hypothesis that there exist a key person who control the conversation. We assume all participants other than the key person pay their attention to the key person in usual case. We call the key person *focusee* in this sense. We also assume this key person owing the right to speak next (except the interrupting utterances, which can be uttered any participants). We call the key person *turn holder* in this sense. The focusee and the turn holder are the different title of the same key person. The focusee or the turn holder is not fixed but changing dynamically in the development of the conversation.

It is very important for the robot to pay his attention to the focusee.

It is desired for the robot to share the common interest with other participants. It is unnatural for other participants if the robot cannot pay attention to the person to whom all other participants paying attention.

Paying attention to the focusee has another meaning. As we mention before, the focusee is equivalent to the turn holder. The turn holder is expected to speak next. If the robot focusing on the person other than turn holder, then the turn holder can not speak to the robot. In this case, the rhythm of the con-

versation falls into disorder. The robot may miss a chance to get message. You may think it is enough for the robot to begin to focus on the speaker after the speaker started to speak, but it is too late to realize natural conversation. Therefore, the ability of the robot to focus timely on the focusee/turn holder be indispensable for multi-person conversation.

To realize this function, the face direction recognition is utilized. When we pay attention to a person, then we face him. So face direction is the good queue to identify the focusee.

Acting as the focusee/turn holder is also very important issue in the multi-person conversation. Robot will be the focusee if he succeed to receive attention from the other participants. Robot can focus on other participants freely, while he is the focusee.

However, there exist a participant who tries to get the turn by interruption. The robot has to decide whether he gives the turn to the interrupter or he keeps the turn. In case of keeping the turn, the robot has to appeal the interrupter that he is not ignoring but he has no wish to give the turn to the interrupter now. In case of accepting the interruption, the robot has to give the turn. To realize this function, we prepared facial expression (described in section 3.6). The robot expresses his will to the interrupter using his words and his gaze. To maintaining the turn, the robot say "Just moment please" with a side-glance. To accept the interruption, the robot faces to the interrupter and gives him the turn.

3 ORGANIZATION AND COMPONENT OF THE ROBOT

In this section, we describe the component technologies in the robot and the organization structure of them, and show how the functions mentioned above section are realized.

3.1 Hardware and computer network

The robot has two compact color CCD cameras on his head and 2 d.o.f. for each of the camera. These cameras are used to show his attention as well as to capture the image. Head, on which the cameras are mounted, has 2 d.o.f. on the bottom of it (namely neck part). Two microphones are mounted on the side of the head. Two 4 d.o.f. arms are mounted on the side of the body. Under the body he has 1 d.o.f. for waist, and wheels to move around.

We mounted two computers on the robot. One is for the control of the entire servo motors on every d.o.f.. One is for the speech synthesis. On the outside of the robot, we mounted 7 workstations to process sound & image information. All workstations and robot are connected by Ethernet network.

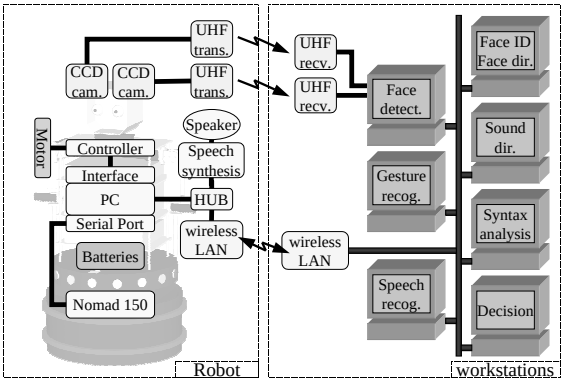

Figure 3: Hardware of the robot and computer network outside

3.2 Detection of sound source

To detect who is speaking, our robot will detect the sound source.

Sound input of the far side microphone from the speaker will affected by the complicated transfer function caused by the robot's head, while sound input of near side microphone to the speaker will have no effect. Head transfer function is dependent on the sound direction. Therefore, the spectral difference of the two microphones shows typical pattern according to the sound direction. Using this feature, we estimated approximate sound direction.

3.3 Face detection and recognition

To recognize the intended receiver of the message, or to recognize the focusee of the each participant, the robot utilizes the face direction of the participants. To recognize who is the speaker, the robot utilizes the face recognition.

Firstly, we find out a face region in captured image using the color and the angle information (We assume color of the face is Japanese skin color and the robot and speaker's head is at the same height) . Then, the feature vectors of the face region is extracted using eigenface method. These features are applied to the Bayesian classifier, person identification and face direction identification are performed.

The category number of the face recognition is 12 (people in our laboratory). The resolution of the face direction recognition is 30 degree.

3.4 Gesture recognition

To recognize the demonstrative action in the space sharing, the robot utilizes the gesture recognition.

From the captured image, the positions of head, hands, shoulders and elbows are extracted. In this

process, the color and the edge information are utilized. Using relative positions of these 7 points, we can decide the posture of the person.

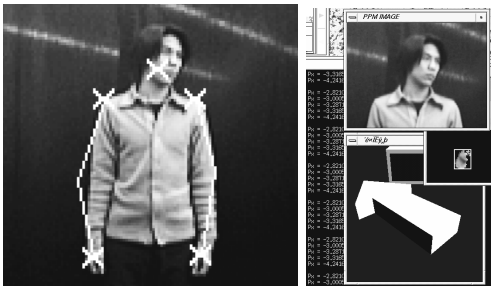

Figure 4: Detection of face direction and hand position

3.5 Speech recognition

System can recognize continuously uttered Japanese sentences. The bigram based frame synchronous decoder constructs the recognizer. Vocabulary is selected to cover the commands to the robot, the questions about the functions of the robot, and the questions about the personal information of the member of our laboratory. Vocabulary size is about 1000.

3.6 Gestural expression

To share the common space information and to clarify the turn holder, face direction control plays a very important role. As we mentioned before, the robot has 2 d.o.f. in his eyes and 2 d.o.f. in his neck. These redundant d.o.f. can be utilized for several expression of his will. In case that the robot squarely looks at a person, the robot can express his will to give the turn to him. While, in case that the robot gives a side-glance to a person, the robot expresses his will to keep the turn.

To appeal the intended receiver of the message, the robot uses the gaze control. The robot will gaze his eyes to the participant whom he wants to talk with. The participant who gets eye contact can perceive the intention of the robot. Other participants who failed to get eye contact can also know with whom the robot intended to speak.

Face direction and hand gestures are utilized to share the common space information. The robot dose demonstrative action with his arms and face to the intended place. By these actions, the robot shares the space information with participants.

4 BEHAVIOR OF THE ROBOT

In this section, we introduce the behavior of the robot.

The robot can answer the question about him and position information of other person. The robot will not response to sound input, while user is not looking to him and seek the person who is the turn holder by other person's gaze. When person is looking at him, and asking some question, then he will speak answer to person, at the same time gaze to him. Robot response to the user by pointing action, when question is about the position of other person.

Figure 5 shows the example of the behavior of the robot.

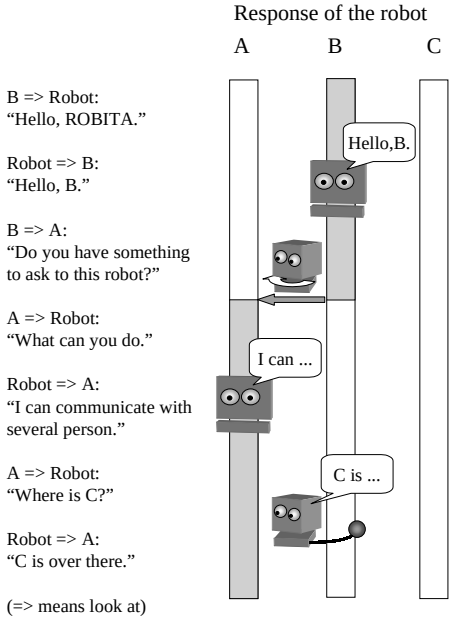

Figure 5: Words and response of the robot

5 CONCLUSION

In this paper, we clarified the characteristics of the multi-person conversation and introduced our multi-model robot designed for the multi-person conversation. We believe that information input from image and sound, and output of the system in non-verbal form is essential for conversation system for multi-person.

References

[1] H.Kikuchi, M.Yokoyama, K.Hoashi, Y.Hidaki, T.Kobayashi, K.Shirai (1998), Controlling Gaze of Humanoid in Communication with Human, *Proc. of Intl. Conference on Intelligent Robots and Systems*, pp.255-260