特集論文

Dynamic Communication of Humanoid Robot with Multiple People Based on Interaction Distance

Tsuyoshi Tasaki — Graduate School of Informatics, Kyoto University
tasaki@kuis.kyoto-u.ac.jp, http://winnie.kuis.kyoto-u.ac.jp/~tasaki/

Shohei Matsumoto — (affiliation as previous author)
shohei_m@kuis.kyoto-u.ac.jp, http://winnie.kuis.kyoto-u.ac.jp/~shohei_m/

Hayato Ohba — Faculty of Engineering, Kyoto University
hayato@kuis.kyoto-u.ac.jp, http://winnie.kuis.kyoto-u.ac.jp/~hayato/

Shunichi Yamamoto — Graduate School of Informatics, Kyoto University
shunichi@kuis.kyoto-u.ac.jp, http://winnie.kuis.kyoto-u.ac.jp/~shunichi/

Mitsuhiko Toda — NTT Docomo, Kansai,
mtoda@kuis.kyoto-u.ac.jp, http://winnie.kuis.kyoto-u.ac.jp/~mtoda/

Kazunori Komatani — Graduate School of Informatics, Kyoto University
komatani@kuis.kyoto-u.ac.jp, http://winnie.kuis.kyoto-u.ac.jp/~komatani/

Tetsuya Ogata — (affiliation as previous author)
ogata@kuis.kyoto-u.ac.jp, http://winnie.kuis.kyoto-u.ac.jp/~ogata/

Hiroshi G. Okuno — (affiliation as previous author)
okuno@kuis.kyoto-u.ac.jp, http://winnie.kuis.kyoto-u.ac.jp/~okuno/

keywords: proxemics, distance based behavoir selection, subsumption architecture, interaction, humanoid

Summary

Research on human-robot interaction is getting an increasing amount of attention. Since most research has dealt with communication between one robot and one person, quite few researchers have studied communication between a robot and multiple people. This paper presents a method that enables robots to communicate with multiple people using the "selection priority of the interactive partner" based on the concept of *Proxemics*. In this method, a robot changes active sensory-motor modalities based on the *interaction distance* between itself and a person. Our method was implemented into a humanoid robot, *SIG2*. *SIG2* has various sensory-motor modalities to interact with humans. A demonstration of *SIG2* showed that our method selected an appropriate interaction partner during interaction with multiple people.

1. Introduction

Studies of human interaction with the robots reported by *Robita* [Matsusaka 99], *Robisuke* [Fujie 04], *SIG* and *SIG2* [Okuno 03, Okuno 04], *ASIMO* [Sakagami 02], *AIBO* [Kaplan 04], *Robovie* [Ishiguro 02], *Kismet* [Breazeal 01], *Wamoeba* [Ogata 00], and *WE-4* [Miwa 03] have obtained much attention. Since almost all of them have dealt with only one-to-one communication between a robot and a person, quite few researchers have studied the methodology for communications between a robot and multiple people who are willing to communicate with the robot at the same time. If a human support robot is going to be developed, robots must be able to interact effectively with multiple people at the same time. This paper presents a design method for such human-robot communications.

The *distance* between a robot and each person is one of the most important issues in interaction with multiple people. Basically, a robot cannot communicate with multiple people at the same time, except when the people can be regarded as one unit, such as an audience at a lecture. Robots may select an "interactive partner" dynamically, based on various criteria such as "intimacy". They may also change

their sensory devices and behavior according to the situation. If people talk to the robot far away from the robot, their sound level is low and thus it is difficult to separate speech from a mixture of utterances. If the robot speaks to distant people, it may cause them to misunderstand that the robot can hear them at the distance. To avoid such misunderstandings, it should not utter voices but use gestures. On the other hand, if people are very close to the robot, such a burden may be alleviated, that is it should speak. In addition, tactile sensors, such as skin sensors, may be used.

From the viewpoint of the *interaction distance*, appropriate behaviors and sensory devices should be selected. We call this kind of modality *sensory-motor modality* for human-robot interaction. Main issues in selection of appropriate modality are listed below:

- selection of appropriate behaviors,
- selection of appropriate sensors, and
- focus of attention.

Robita is a conversation robot that can participate in group discussion [Matsusaka 99]. Two people sitting on a chair interact with each other and *Robita*. During interaction, *Robita* obtained auditory inputs through a headset microphone worn by each participant. In this sense, its interaction model did not depend on the interaction distance and used the fixed sensory-motor modality.

Robita maintains various kinds of information on the blackboard and selects an appropriate module with highest priority [Kim 02]. The system architecture is based on inter-module cooperation consisting of priority management, situated observation, and data exhibition/message dispatch systems. The priority of module is given in advance by a task designer (software developer). Therefore, *Robita*'s behavior is a result of priority based execution control.

Robovie [Kanda 04] is a social robot and its field-test had been carries out at an elementary school for two months. *Robovie* successfully understood friendly relationship among students in front of it by tracking their RFID tags attached to student's name tag. Since this field-test needed a precise identification of students, only RFID tag information was utilized. Although *Robovie* did not exploit physical distance information, it extensively utilized mental distance information.

Complex emotional behaviors were realized on *WE-4* (Waseda Eye, No.4) [Miwa 03] by focusing on the distance in the mental space. Emotional states were represented by a mood vector and inputs from visual, tactile, auditory, and olfactory sensors were associated to some mental state. The temporal decay of mood was also represented by an equation. Although sensory input was simple, *WE-4* successfully produced emotional expressions.

SIG2 [Okuno 03, Okuno 04] tracks multiple people who are either talking or not talking by integrating visual and auditory localization. It could perform various kinds of visual and auditory scene analyses including face localization and recognition, sound source localization and separation, and automatic speech recognition. Although it has various sensory-motor modalities, its behaviors are only passive; it can track and turn toward a speaker.

In this paper, we design a method of human-robot dynamic communication in which the robot selects an interactive partner from multiple people by assigning "priority" based on the interaction distance. In this method, the robot refines its recognition and behavior by selecting appropriate sensory-motor modalities based on the interaction distance.

In Section 2, "Proxemics", a sociological theory is introduced as the basic concept of our method and the details of our method are described. In Section 3, a humanoid robot used in this study and actual implementation are described. In Section 4, some demonstrations of the robot's behavior when communicating with multiple people are described. In Section 5, the effects of the proposed method are discussed. Section 6 concludes this paper.

2. Communication based on Interaction Distance

In this section, we set up a communication methodology by using *Proxemics* [Hall 66]. The interaction distance between a robot and people in communication from viewpoints of sociology, in particular, Proxemics, and sensory capabilities. The former analyzes the spatial relationships on the basis of communication between persons. The latter analyzes the performance of recognition by various sensors of a robot based on the distance to an object. We map Proxemics as a general theory to a specific platform, *SIG2* humanoid robot by using its sensory capabilities.

2·1 Proxemics

In sociology studying man-to-man communication, Proxemics describes the role of distance, or spatial

relationships that can impede or promote the act of communication. The spatial territory for interpersonal communication is classified into four groups:

- *Intimate distance* (approx. 50 cm or less) — people can communicate via physical interaction and express strong emotion by embracing or whispering.
- *Personal distance* (approx. 50–120 cm) — people can make conversations among good friends.
- *Social distance* (approx. 120–360 cm) — people can make conversations among acquaintances.
- *Public distance* (approx. 360 cm or more) — people who have no personal relationships with each other can comfortably coexist.

The distance values shown in parentheses are just typical examples and they usually depend on a person's personality and cultural backgrounds. For a robot, we assume that a set of the effective distance of each sensory capability define an instance of the distance values for four categories.

2·2 Effective distances for sensors and devices

Since most sensors and devices are not effective for all ranges, we assess effective distance for them. Effective distance is a kind of constraints for sensory-motor modalities. It will reduce the complexity in designing interaction systems.

As input sensors, tactile sensors are effective within the reach of people, that is, intimate distance. If a robot knows that a target person is at intimate distance, it may focus its attention on him/her and use either speech recognition or face recognition for interaction. As output devices, normal loud speakers are not appropriate for interactions at public distance, because they deliver sounds to all the people around the robot. A sound spotlight based on a parametric loud speaker may be useful to deliver sounds to the people in a particular direction.

The platform robot, *SIG2*, uses tactile sensor, face recognition, sound source localization, sound source separation and recognition. We will assess each sensory capability from the viewpoint of the interaction distance or Proxemics in the next section.

2·3 Robot Intimacy based on Proxemics

Another factor in determining behaviors is *intimacy*. Proxemics suggests that the more intimate the communication, the nearer the target person stands. The parameter of intimacy is introduced to reflect the relationship between a robot and humans. The robot uses this parameter to determine communication priority among multiple people in a situation, and then behaves according to its relationship with each person.

The parameter of "Intimacy", I, ranges from 0 to 1. It represents the intimacy of the relationship between a robot and a human. Since I changes dynamically during communication, its level changes according to the following equations:

$$I(0) = P, \tag{1}$$

$$\frac{dI}{dt} = \left(\frac{I+P}{2}\right) \cdot D - I \cdot \left(\frac{P \cdot I + 1}{2}\right) + S_k. \tag{2}$$

The term P is a constant parameter defined a priori as the robot personality. The first term on the right-hand side of Equation (2) shows the influence of the distance. The parameter of the distance, D, is defined as follows:

$$D = \begin{cases} 0.04 & \textit{if} \quad \text{Intimate distance}, \\ 0.02 & \textit{if} \quad \text{Personal distance}, \\ 0.0 & \textit{otherwise}. \end{cases} \tag{3}$$

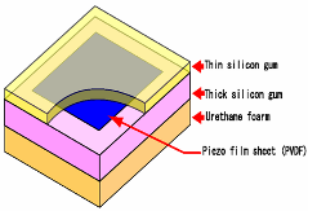
The term I is defined as the summation of the friendliness of the robot and intimacy of its relationship with a person. If the robot recognizes the person as someone it is intimate with, I increases. If the robot recognizes the perosn as someone it is not intimate with, I decreases. The second term of Equation (2) is a damping factor. If the robot has no communication with people for a while, the I converges to 0. The S_k is a parameter of the influence of stimulus. It changes I based on the human behavior.

3. Humanoid *SIG2* and Its Capabilities

The platform we used was the humanoid robot, *SIG2*, shown in Figure 1. *SIG2* has one omni-directional microphone on each side of its head. Each microphone is embedded in the eardrum of a model of a human outer ear made of silicon (Figure 1-b). Its head and upper body are covered with soft skin-like material containing 19 patches of tactile sensors (Figure 1-c). A directional parametric speaker is located at its waist (Figure 1-d). It generates a sound beam of about 20° up to 10 meters. *SIG2* also has an omni-directional (normal) loud speaker, of which effective range is up to about 5 meters.

(a) *SIG2*

(b) ear



(c) tactile sensor

(d) directional speaker

Fig. 1 *SIG2* and its parts; (b) ear, (c) piezo tactile sensor, and (d) directional loud speaker.

In the remaining of this section, each sensor is explained and its effective range for each sensor is assessed from the viewpoint of Proxemics. The results are summarized in Table 1.

3·1 Tactile Sensors

Each tactile sensor, which consists of piezo elements covered by silicon, can detect the pressure velocity of its patch. It can recognize three kinds of contact: *touch*, *rub*, and *hit*. Velocity versus time for a hit and a rub are shown in Figure 2.

The tactile sensors are effective within where people can touch *SIG2*. The average length of adult's arm is about 70cm, and thus the effectuve distance for tactile sensor is up to 50cm. This distance is similar to the intimate distance.

3·2 Face Localization and Recognition

SIG2 can measure the distance to its partner using stereovision which uses two cameras in its head. Since its visual processing detects multiple faces, extracts, identifies, and tracks each face simultaneously, the size, direction and brightness of each face changes frequently. We use MPIsearch [Fasel 02] to attain robust face detection, as shown in Figure 3.

After an extracted face is identified, it is projected into discrimination space, and its distance, d, from each registered face is calculated [Okuno 03]. Since this distance depends on the degree (L, the number of registered faces) of the discrimination space, it is converted to a parameter-independent probability, P_v:

$$P_v = \int_{\frac{d^2}{2}}^{\infty} e^{-t} t^{\frac{L}{2}-1} dt. \quad (4)$$

A discrimination matrix is created in advance or on demand by using a set of variations of the face with an ID (name). This analysis is done using online linear discriminant analysis.

Face localization of *SIG2* uses MPIsearch, which requires at least an image of 12 by 12 pixels to detect a face. The camera of *SIG2* provides such images at the distance of 4 to 5 meters. In general, the effective distance of face localization is up to public distance, but the performance of face localization is strongly influenced by lighting conditions.

To reduce ambiguities of face localization due to lighting conditions and occlusion and improve sound source localization, face localization and sound localization are integrated [Nakadai 04].

3·3 Sound Source Localization

Sound source localization is performed analogously to human perception; *SIG2* uses the two microphones embedded in its head. We used the sound source localization and separation system, the Active Direction-

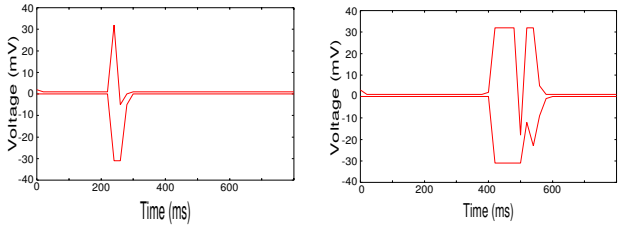


Fig. 2 Responses of tactile sensor, *hit* (left) and *rub* (right).

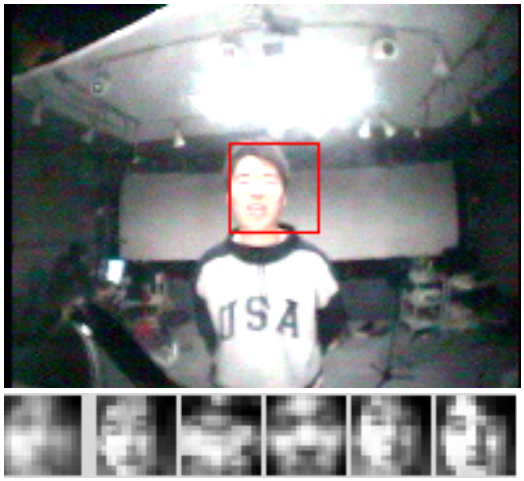


Fig. 3 Face Localization and Recognition.

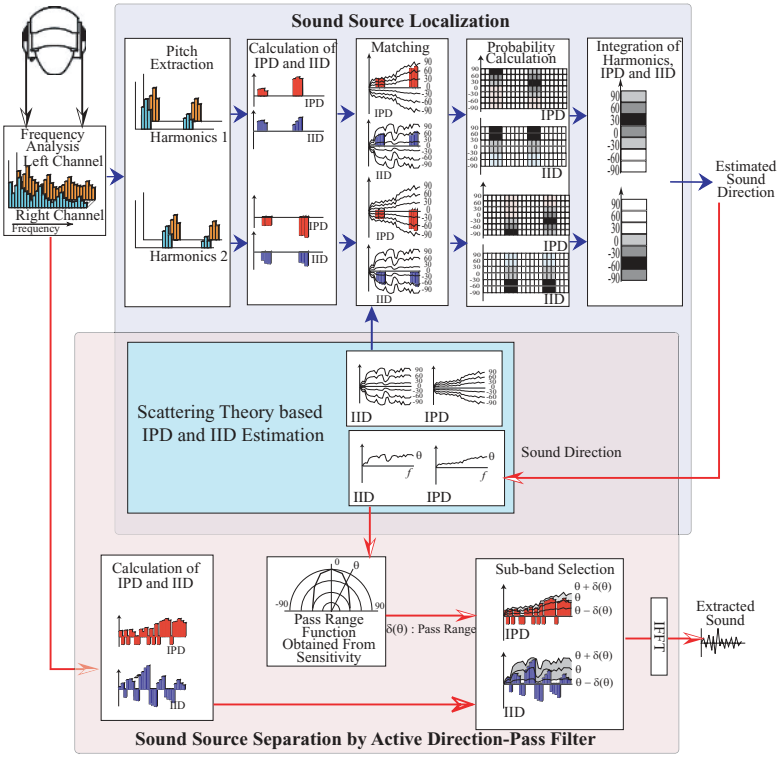
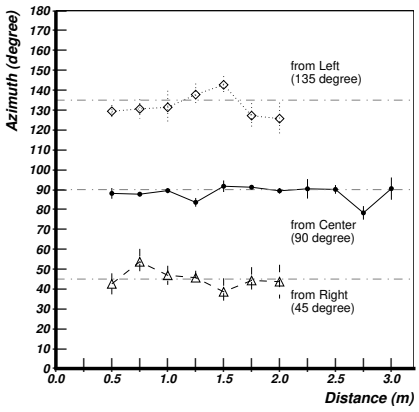
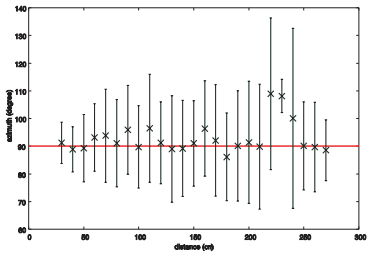


Fig. 4 Sound Source Localization and Separation System.



(a) Our laboratory (right, center, left)



(b) Kyoto University Museum (center)

Fig. 5 Sound source localization at various distances.

Pass Filter (ADPF) [Nakadai 02] specified in Figure 4).

To localize sound sources with two microphones, first a set of peaks are extracted for the left and right channels. Then, identical or similar peaks of left and right channels are identified as pairs and each pair is used to calculate interaural phase difference (IPD) and interaural intensity difference (IID). IPD is calculated from frequencies of less than 1500 Hz, while IID is from frequency of more than 1500 Hz. Because auditory and visual tracking involves motor movements, which cause motor and mechanical noises, ADPF uses heuristics to reduce internal burst noises caused by motor movements.

From IPD and IID, the epipolar geometry with scattering theory is used to obtain the direction of sound source [Nakadai 04]. The key ideas of their real-time active audition system are twofold; one is to exploit the property of the harmonic structure (fundamental frequency, $F0$, and its overtones) to find a more accurate pair of peaks in left and right channels. The other is to search for the direction of the sound source by combining the belief factors of IPD and IID using the Dempster-Shafer theory.

The effective distance of sound source localization in average and standard deviations was estimated for two rooms; our laboratory (Figure 5(a)) and the Entrance Hall of Kyoto University Museum (Figure 5(b)). The size of our laboratory is about 7 m×5 m and its reverberation time (RT20) is about 0.3 to 0.4 sec. The reverberation of the latter is much stronger than the former, because the walls, floor and ceiling are covered with concrete and its walls are made of glass. The data for localization is a single voiced sound saying “oh-i”. For our laboratory, three directions are evaluated separately. The horizontal direction (azimuth) is specified from right (0°) to left (180°), and the center is 90°.

The errors of localization remain small for our laboratory, while they gradually increase as the distance is longer. Therefore, this assessment confirms that the sound source localization remains stable up to around 3 meters except at some distances. Reverberation and reflections caused by the room acoustical conditions affect the performance of sound source localization. The results of Kyoto University Museum showed the strong influence of reflection by the glass wall. Nevertheless, the sound source localization is effective up to the public distance. As mentioned in the previous subsection, ambiguities in sound source localization may be reduced by integrating face localization if available.

3·4 Sound Source Separation

The architecture of the ADPF is shown in lower dark area in Figure 4. The ADPF separates out sound sources from a spectrum of input sound, IPD and IID of the input sound, and sound source direction. The details of the ADPF algorithm are as follows:

(1) The pass range, $\delta(\theta_s)$, of the ADPF is selected according to pass range function, δ. Its minimum value is straight in front of *SIG2*, because the ADPF has its maximum sensitivity there. The function, δ, has a larger value at the periphery because of a lower sensitivity. Let us $\theta_l = \theta_s - \delta(\theta_s)$ and $\theta_h = \theta_s + \delta(\theta_s)$.

(2) From a sound direction, the IPD, $\Delta\varphi_E(\theta)$, and IID, $\Delta\rho_E(\theta)$, are estimated for each sub-band using auditory epipolar geometry. Likewise, the IPD $\Delta\varphi_H(\theta)$ and IID $\Delta\rho_H(\theta)$ are obtained from HRTFs.

(3) The sub-bands are collected if the IPD and IID satisfy the specified condition.

$$f \le f_{th} : \Delta\varphi_s(\theta_l, f) \le \Delta\varphi(f) \le \Delta\varphi_s(\theta_h, f), \text{and}$$
$$f > f_{th} : \Delta\rho_s(\theta_l, f) \le \Delta\rho(f) \le \Delta\rho_s(\theta_h, f). \quad (5)$$

(4) A wave consisting of collected sub-bands is constructed.

3·5 Speech Recognition for Separated Sound

Making speech recognition robust against noises is one of the hottest topics in the speech community. Approaches have been developed, such as multi-condition training and missing data [Barker 01, Renevey 01], that are, to some extent, efficient at recognizing speech with noise. However, these methods are of less use when the signal to noise ratio is as low as 0 dB as occurs with a mixture of speech from different voices. In this case, speech enhancement by a front-end processing is necessary. This kind of speech enhancement is efficient for speech recognition in higher signal-noise ratio, though such approach has not been studied so much. Then, we propose speech recognition using multiple acoustic models to use the sound source separation by the ADPF as front-end processing.

The Japanese automatic speech recognition software "Julian" was used for automatic speech recognition (ASR). For three simultaneous talkers, acoustic models were created for each talker at every 10° from -90° to 90°. Because we used 17 directions and three speakers, 51 training datasets were obtained. In speech recognition, 51 ASRs were processed against an input in parallel, and each ASR used a different acoustic model. Then the system integrated all the results of ASRs and output the most reliable result. The average rate of isolated word recognition for three simultaneous talkers at 1 meter with 30°, 90°, and 150° was about 80% [Nakadai 04].

The effective distance of automatic speech recognition is estimated for an ideal case where people talk alternatively and do not talk simultaneously. A loud speaker was located at the center (90°) at every 50 cm from 50 cm to 3 meters. 200 words of ATR phonetically balanced corpus are played. Speaker-independent acoustic model for center-direction was created by using the data of non-target talkers. The result of isolated word recognition is shown in Figure 6. The performance is deteriorated for a longer distance. At 2.5 meters, the localization was poor as is shown in Figure 5(a) and thus the performance of recognition was also poor. Automatic speech recognition is considered effective up to for up to around 1.5 meters.

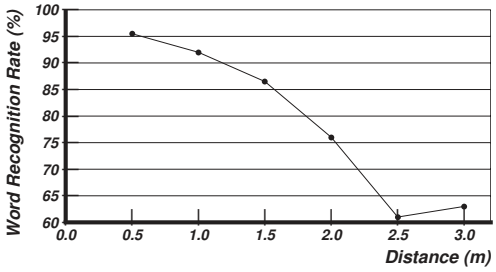


Fig. 6 Isolate word recognition at various distances.

4. Design of Interaction based on Proxemics

4·1 Mapping Proxemics to *SIG2*

Based on the observations of the previous section, the effective distances for input sensors and output devices in terms of Proxemics are summarized in Table 1. The actual selections are as follows:

- Public and social distances — *SIG2* can locate humans using skin color information of the vision system and locate sound sources. Both localization systems are integrated at the sensing module levels to reduce ambiguities of localization.
- Personal distance — Besides the functions mentioned above, *SIG2* carries out sound source separation, speech recognition, and face recognition.
- Intimate distance — Besides the functions used in calculating personal distance, *SIG2* recognizes

Table 1 Summary of effective distances for sensors and output devices.

Modality	Input sensors						Output devices		
Proxemics *Distance*	Skin sensor	Skin color (face local.)	Sound localization	Sound separation	Speech recognition	Face recognition	Omni-direc. speaker	Sound spotlight	Tracking· gesture
Intimate	√	√	√	√	√	√	√		√
Personal		√	√	√	√	√	√		√
Social		√	√				√	√	√
Public		√	√				√	√	√

three kinds of contact: *touch*, *rub*, and *hit*.

SIG2 has a four-degree-of-freedom rotation such as nod, incline, rotation of its neck and rotation of its body, movement using its cart, and utterance enabled by the two kind of speakers (directional and omni-directional).

Based on the distance to a person, *SIG2* selects movement functions:

- Intimate distance — *SIG2* uses the omni-directional (normal) loud speaker for utterances.
- Personal distance — Besides using the omni-directional loud speaker for utterances, speakers are tracked and gestures are facilitated by four motors
- Social distance — Besides the functions used in personal distance, the directional loud speaker is used to talk to a person standing far away from *SIG2*.
- Public distance — Besides the functions used in social distance, *SIG2* can use the cart to get close to the target person or people.

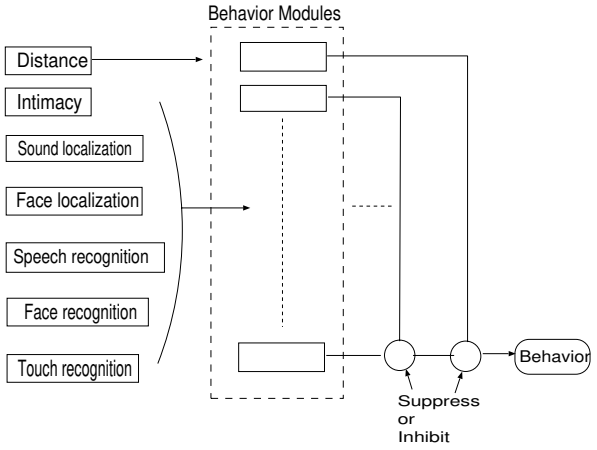


Fig. 7 Conceptual overview of behavoir selection system.

4·2 Implementation using Subsumption Architecture

Our method dynamically determines the priority of various modalities of input sensors and output devices based on the interaction distance. Since all the modalities are not required to implement a particular behavior, a subsumption architecture (*SA*, hereafter) [Brooks 86] is used to design the behavior selection system. In other words, the priority relations are embedded in *SA*.

Robita also uses priorities for modules which are determined manually for a particular situation or task and determined a behavior by selecting the highest priority of competing modules. *Robovie* used situated modules (scenarios) for various situations and determined a behavior by calculating the connection weights of modules. Our system is not superior in scalability to these two systems, but easier to implement and showed various interesting behaviors [Okuno 02].

The actual system is implemented by distributed processing. It consists of 6 Linux PCs connected by Fast Ethernet; one for visual processing, one for auditory processing, one for tactile processing, one for motor control, one for behavior selection and one for sound generation. This paper focuses on the behavior selection system, because it is a new module since [Nakadai 02].

The concept of behavior selection system is depicted in Figure 7. A sensory event represented in symbolic data is given to a behavior module if it is needed. When a behavior module receives sensory events, it starts processing and outputs a candidate of behavior. Such active behavior modules run concurrently under the behavior selection system. In addition, distance and intimacy modules play a role of metacontroller to controls active behavior modules from a higher level.

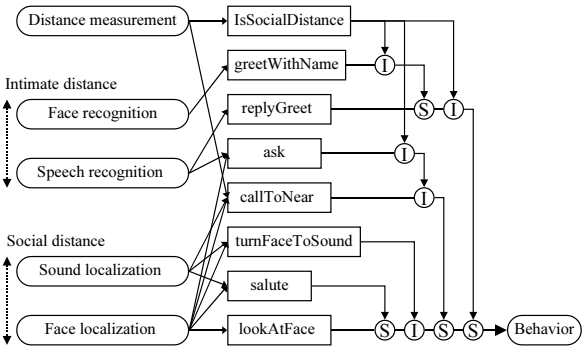


Fig. 8 Modules implemented for Scenario 1.

Step 1 **Step 2** **Step 3**

Fig. 9 Shapshots of Scenario 1.

Distance module maintains a 2D-map of objects by integrating sound source localization, face localization and stereo vision [Nakadai 02]. The output of upper modules suppresses or inhibits that of lower modules to subsume the output of behavior modules.

5. Experiments

We conducted two experiments to confirm the validity of interaction based on the interaction distance. The outlines of these experiments follow below.

5·1 Scenario 1: Selecting sensory modalities based on distance

In this experiment, *SIG2* interacted with two people who spoke from different distances, far and near, by changing input modalities. *SIG2* urged the farthest person from itself to approach. The structure of the *SA* used in this section is shown in Figure 8.

Step 1 Person A said "Hello, SIG2." at a social distance.

After localizing the sound, *SIG2* turned to Person A (turnFaceToSound), and after localizing the face, it continued looking at him (lookAtFace). It detected that he was positioned at a social distance by using the stereovision. Consequently, calling Person A's name (greetWithName) and replying with a greeting (replyGreet) were inhibited.

Step 2 Person B approached with in an intimate distance and said "Hello SIG2."

After localizing the sound and face, *SIG2* turned to Person B and continued looking at him. Because Person B was positioned at an intimate distance, *SIG2* bowed slightly (item salute), called Person B's name, and greeted Person B.

Step 3 Person A called to *SIG2*.

After localizaing the sound and face, *SIG2* turned to Person A and continued looking at him. Greeting Person A was inhibited by the history of Person A's behavior. *SIG2* requested that Person A approach (callToNear). Asking about Person A's business (ask) became active but was is inhibited because of the social distance.

Step 4 Person A followed instructions and approached *SIG2*.

After localizing the face, *SIG2* continued looking at Person A. *SIG2* detected Person A was positioned at an intimate distance. Then it asked Person A's business. Requesting that Person A approach was inhibited.

5·2 Scenario 2: Changing behavior based on intimacy

In this experiment, *SIG2*, between two people, changed its conversation partner based on intimacy. The *SA* used in this section is shown in Figure 10.

Step 1 Person A greeted within an intimate distance.

After localizaing the sound and face, *SIG2* turned to Person A and continued looking at him. *SIG2* bowed slightly, called Person A's name, and replied to Person A because of their intimate distance. Its intimacy with Person A increased.

Step 2 Person B greeted *SIG2* from a social distance.

After localizaing the sound and face, *SIG2* turned to Person B and continued looking at him. Calling his name and offering a greeting was inhibited because of the social distance. *SIG2* compared the intimacy it experienced with Persons A and B, and then returned to Person A, with whom it had higher intimacy (turnToIntimatePerson).

Step 3 Person A rubbed *SIG2*.

The intimacy with Person A increased (updateIntimacy).

Step 4 Person B called to *SIG2* from a social distance.

After localizaing the sound and face, *SIG2* turned

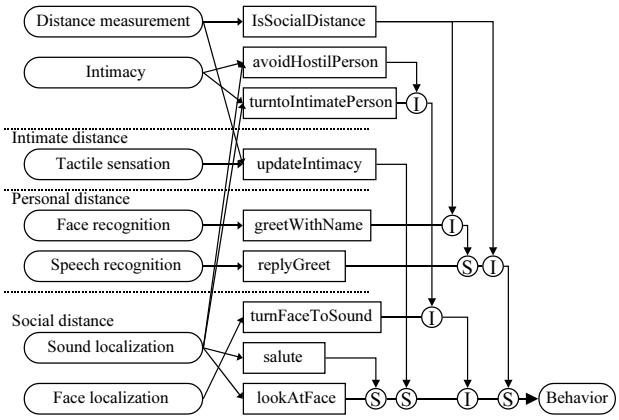


Fig. 10 Modules implemented for Scenario 2.

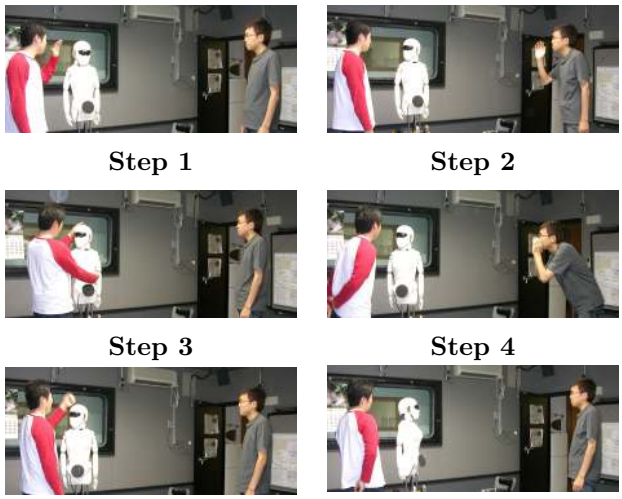


Fig. 11 Snapshots of Scenario 2

to Person B with increasing frequency. However, *SIG2* continued looking at Person A because the intimacy with Person A was over the threshold value (turnToIntimatePerson). *SIG2* did not reply to Person B.

Step 5 Person A hit *SIG2*.

The intimacy with Person A decreased, and *SIG2* began avoiding him (avoidHostilePerson).

6. Discussion and Further Work

In this section, we discuss about oservations obtained in this paper and future work.

6·1 Sensory-motor modalities

In this paper, we presented the selection method of appropirate sensory and motor devices, that is, selection of sensory-motor modalities. As a robot has more sensory-motor modalities, their appropriate classification and usage of appropriate modalitles are mandatory. By using sumsumption architecture with the interaction distance and intimacy as presented in this paper, a robot automatically selects such modalities. Therefore, priorities of sensory-motor modalities are embedded in the subsumption architecture with the interaction distance.

We have developed another interaction system, a game playing system, by integrating speech recognition and speech generation as well as tactile sensors and gestures. This game playing system is also implemented by *SA* with sensory-motor modalities. Therefore, the development of this system wss easy and rapid.

6·2 Effective distance of sensory-motor modalities

We analyzed the effective distance for each sensory-motor modality and obtaned a general guideline according to the interaction distance. In addition, the Proxemics based distances are mapped to those for a particular robot by using the effective distance of its sensory-motor modalities.

The effective distances may be affected by the environments, in particular, lighting conditions or acoustic environments. The future work includes automatic adaptation of the effective distances against environmental changes.

6·3 Representation of effective distance

In this paper, we use a simple 2D-map to represent objects and interaction partners. To persue more sophicsticated representation, we are currently developing a new spatial map of proxemics and the effective distance of sensory-motor modalities. The new itneraction system is under development and field tests had been carried out at Kyoto University Museum for two months. We are currently ananlysing the subjective impression, which will be submitted as a separate paper in a near future.

6·4 Evaluation methodology

We have realized that the method presented in this paper works well for implementating new interaction systems as well as real-time systems. However, such an evaluation is qualitative, but not quantitative.

The future work may include establishing evaluation methodologies. On the contrary, we may give up such common evaluating methodologies and accept a

robot behavior selection systems as is. If the task is fixed, it may be easier to compare between tasks. If the task is social or a human-partner, there are no abosolute measures for evaluation. This kind of disputes will remain for ever.

7. Conclusion

We presented a model of dynamic communication for a humanoid robot to interact with multiple people. The robot selects an appropriate interaction partner based on the interaction distance. Proxemics classifies the interaction distance into four categories and the effective distance of robot's sensory-motor modalities specifies the actual range of each Proxemics distance for the robot. Subsumption architecture is used to design behavior selection system and two scenarios are demonstrated to confirm that the method presented in this paper works well.

Acknowledgments

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research in Informatics, the JPSP 21st Century COE Program on informatics research for development of knowledge society infrastructure, and Artificial Intelligence Research Promotion Foundation. The original *SIG2* was developed by JST Kitano Symbiotic Systems Project. The authors thank Dr. Kazuhiro Nakadai of HRI-Japan and Dr. Hiroaki Kitano of JST Kitano Project for their collaborations.

♢ References ♢

[Barker 01] J. Barker, M. Cooke, and P. Green: Robust asr based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. *Proc. of European Conference on Speech Communication Technology (EUROSPEECH-2001)*, 213–216, 2001.

[Breazeal 01] C.L. Breazeal: *Designing Sociable Robots*, *A Bradford Book*, 2001, ISBN 0262025108.

[Brooks 86] R.A. Brooks: A Robust Layered Control System For A Mobile Robot, *IEEE Journal of Robotics and Automation*, pp.14–23, Vol.2, No.1, 1986.

[Dempster 67] A. Dempster: Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[Fasel 02] I. Fasel, and J.R. Movellan: Comparison of neurally inspired face detection algorithms, . UAM, 2002. *Proc. of International Conference on Artificial Neural Networks (ICANN 2002)*, 1395–1401. 2002.

[Fujie 04] S. Fujie, Y. Ejiri, K. Nakajima, Y. Matsusakai, and S. Kuota: A Conversation Robot Using Head Gesture Recognition as Para-Linguistic Information, *Proc. of IEEE International Workshop on Robot and Human Communication (Ro-Man 2004)*, 159–164, 2004.

[Hall 66] E.T. Hall: *Hidden Dimension*, *Doubleday Publishing*, 1966.

[Ishiguro 02] H. Ishiguro, T. Miyashita, T. Kanda, T. Ono, and M. Imai: Robovie: An interactive humanoid robot, *Video Proc. of IEEE International Conference on Robotics and Automation (ICRA-2002)*, 2002.

[Kanda 04] T. Kanda, R. Sato, N. Saiwaki, and H. Ishiguro: Friendly social robot that understands human's friendly relationships, *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2004)*, 2215–2222, 2004.

[Kaplan 04] F. Kaplan, and V.V. Hafner: The Challenge of Joint Attention, *Proc. of the Fourth International Workshop on Epigenetic Robotics (EpiRobo-2004)*, 67–74, Lund Unviersity Cognitive Studies, 117, 2004.

[Kim 02] K. Kim, Y. Matsusaka, T. Kobayashi: Inter-Module Cooperation Architecture for Interactive Robot, *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2002)*, vol.3, 2286–2291, 2002.

[Matsusaka 99] Y. Matsusaka, T. Tojo, S. Kuota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi: Multi-person Conversation via Multi-modal Interface — A Robot who Communicates with Multi-user, *Proc. of European Conference on Speech Communication Technology (EUROSPEECH-99)*, 1723–1726, 1999.

[Miwa 03] H. Miwa, T. Okuchi, K. Itoh, H. Takanobu, and A. Takanishi: A NewMental Model for Humanoid Robots for Human Friendly Communication – Introduction of Learning System, Mood Vector and Second Order Equations of Emotion –, *Proc. of IEEE International Conference on Robotics and Automation (ICRA-2003)*, 3588–3593, 2003.

[Nakadai 02] K. Nakadai, K. Hidai, H.G. Okuno, and H. Kitano: Real-time speaker localization and speech separation by audio-visual integration, *Proc. of IEEE International Conference on Robotics and Automation (ICRA-2002)*, 1043–1049. IEEE, 2002.

[Nakadai 04] K. Nakadai, D. Matsuura, H.G. Okuno, and H. Tsujino: Improvement of Recognition of Simultaneous Speech Signals Using AV Integration and Scattering Theory for Humanoid Robots, *Speech Communication*, Vol.44 (2004) 97-112, Elsevier.

[Ogata 00] T. Ogata, and S. Sugano: Emotional Communication between Humans and the Autonomous Robot WAMOEBA-2 Which has the Emotional Model, *JSME International Jounal, Series C, Mechanical Systems Machine Elements and Manufacturing*, Vol.43, No.3, pp.568–574, Sept. 2000.

[Okuno 02] H.G. Okuno, K. Nakadai, and H. Kitano: Realizing Audio-Visually triggered ELIZA-like non-verbal Behaviors, *PRICAI 2002: Trendss in Artificial Intelligence*, LNAI 2417, 552–562, Sprinver-Verlag.

[Okuno 03] H.G. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, and H. Kitano: Human-Robot Non-Verbal Interaction Empowered by Real-Time Auditory and Visual Multiple-Talker Tracking, *Advanced Robotics*, Vol.17, No.2 (March, 2003), 115–130, VSP.

[Okuno 04] H.G. Okuno, K. Nakadai, T. Lourens, and H. Kitano: Sound and Visual Tracking for Humanoid Robot, *Applied Intelligence*, Vol.20, No.3 (May/June, 2004), 253–266, Kluwer Publishers.

[Renevey 01] P. Renevey, R. Vetter, and J. Kraus: Robust speech recognition using missing feature theory and vector quantization. In *Proc. of European Conference on Speech Communication Technology (EUROSPEECH-2001)*, 1107–1110. 2001.

[Sakagami 02] Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, and K. Fujimura: The Intelligent ASIMO: System overview and integration, *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2002)*, 2478–2483, 2002.

〔担当委員：間瀬 健二〕

Received November 3, 2004.

Author's Profile

Tsuyoshi Tasaki

He received the B.E. from Department of Information and Mathematical Science, Kyoto University in 2004. He is currently a master course student at Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. He is a student member of IPSJ, RSJ and ISCA.

Shohei Matsumoto

He received the B.E. from Department of Information and Mathematical Science, Kyoto University in 2004. He is currently a master course student at Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. He received IPSJ 67th National Convention Student Award in 2005. He is a student member of IPSJ and RSJ.

Hayato Ohba

He is a undergraduage student at Department of Computer Science and Applied Mathematics, Faculty of Engineering, Kyoto University. He received IPSJ 67th National Convention Student Award in 2005. He is a student member of IPSJ.

Shun'ichi Yamamoto (Student Member)

He recieved the B.E. from Department of Information Science and Mathematical Scientce, Faculty of Engineering, Kyoto University in 2003. He is currently a master course student at Department of Intelligence Science and Technology, Graduate School on Informatics, Kyoto University. A major focus of his research is to build a robust robot audition system in real environments. He received IEEE Robotics and Automation Society Japan Chapter Young Award and IEEE Kansai Chapter Student Award in 2004 and 2005, respectively. He is a student member of IPSJ, RSJ, and IEEE.

Mitsuhiko Toda

He received the B.E. from Department of Information and Mathematical Science, Kyoto University in 2002 and the M.S. from Department of Intelligence Science and Technology, Graduate School of Informatics, the same university in 2004. He received IPSJ 66th National Convention Student Award in 2004. He is currently an engineer, NTT DoCoMo Kansai.

Kazunori Komatani (Member)

Dr. Kazunori Komatani is an Assistant Professor with Graduate School of Informatics, Kyoto University, Japan. He received the B.E. degree in 1998, the M.S. degree in Informatics in 2000, and the Ph.D. degree in 2002, all from Kyoto University. He received the 2002 FIT Young Researcher Award and 2004 IPSJ Yamashita SIG Research Award, both from the Information Processing Society of Japan. His research interests center on spoken dialogue systems. He is a member of IPSJ, NLP, and IEICE.

Tetsuya Ogata (Member)

He received the B.S., M.S., and Dr. of Engineering in Mechanical Engineering in 1993, 1995, and 2000, respectively, from Waseda University. From 1999 to 2001, he was a Research Associate in Waseda University. From 2001 to 2003, he was a Research Scientist in Brain Science Institute, RIKEN. Since 2003, he has been a Faculty Member in Graduate School of Informatics, Kyoto University, where he is currently an Associate Professor. Since 2001, he has been a Visiting Lecturer of the Humanoid Robotics Institute of Waseda University. His research interests include human-robot vocal/sound interaction, dynamics of human-robot mutual adaptation, and multi-modal active sensing with robot system. He received the JSME Medal for Outstanding Paper from the Japan Society of Mechanical Engineers in 2000.

Hiroshi G. Okuno (Member)

He received the B.A. from Department of Pure and Applied Sciences, The University of Tokyo in 1972 and Ph.D from Department of Information Engineering, the same university in 1996. He worked for Nippon Telephone and Telegram Corp., Japan Science and Technology Corp., Science University of Tokyo, and joined Kyoto University in 2001. He is currently a Professor at Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. He is a councilor of JSAI. He received various awards including the 1990 Best Paper Award of JSAI, the Best Paper Award of IEA/AIE-2001, and the 2002 Funai Foundation Scientific Achievement Award. He is a member of IPSJ, JSSST, JCSS, RSJ, ACM, AAAI, ASA, and IEEE.