A Gesture-Centric Android System for Multi-Party Human-Robot Interaction
Yutaka Kondo, Kentaro Takemura, Jun Takamatsu, Tsukasa Ogasawara
Nara Institute of Science and Technology

Natural body gesturing and speech dialogue, is crucial for human-robot interaction (HRI) and human-robot symbiosis. Real interaction is not only with one-to-one communication but also among multiple people. We have therefore developed a system that can adjust gestures and facial expressions based on a speaker’s location or situation for multi-party communication. By extending our already developed real-time gesture planning method, we propose a gesture adjustment suitable for human demand through motion parameterization and gaze motion planning, which allows communication through eye-to-eye contact. We implemented the proposed motion planning method on an android Actroid-SIT and we proposed to use a Key-Value Store to connect the components of our systems. The Key-Value Store is a high-speed and lightweight dictionary database with parallelism and scalability. We conducted multi-party HRI experiments for 1,662 subjects in total. In our HRI system, over 60% of subjects started speaking to the Actroid, and the residence time of their communication also became longer. In addition, we confirmed our system gave humans a more sophisticated impression of the Actroid.

Keywords: human-robot interaction, body gesture, facial expression, multi-party, android

1. Introduction

Human recognition of communication between a person and a robot is essential for successful human-robot interaction (HRI) and human-robot symbiosis (Ishiguro et al., 2001). Most previous research in HRI only dealt with communication between one robot and one person; however, communication is not only one-to-one but can also be multi-party. More recently, robots and computer graphing (CG) agents have been attracting attention as presenters of explanation and guidance for people (Ministry in Japan, 2003). Although several research methods tackled multi-party communication, these methods focused mainly on the recognition of multi-party conversations (e.g., (Nagao & Takeuchi, 1994; Matsusaka, Tojo, & Kobayashi, 2003; Sakamoto, Kanda, Ono, Ishiguro, & Hagita, 2007)).

We therefore have developed an android system that enables multi-party communication through flexible body gesture and facial expression. Gestures such as a pointing are commonly used in all countries as nonverbal communication (Fast, 1970) and are, along with speech dialogue, crucial for HRI. In addition, robots are superior to CG agents as presenters because they can point and...
Kondo et al., A Gesture-Centric Android System for Multi-Party Human-Robot Interaction

Figure 1. Concept of the proposed HRI system that can communicate with multiple people by interruptible communication and gesture adjustment based on a speaker or object location. The purpose of this system is to realize these functions in real-time.

gaze at objects in the real world (Shinozawa, Naya, Yamato, & Kogure, 2005). In our previous study (Kondo, Takemura, Takamatsu, & Ogasawara, 2010), we focused on body gesture planning by especially considering human-likeness and interactivity, which are strongly required for androids to avoid the Uncanny Valley (Mori, 1970; White, McKay, & Pollick, 2007). We have already proposed an online motion planning method named the Reconfigurable Motion Database (RMDB), based on a pre-designed motion database that can reduce a human’s waiting time for communications, while achieving gesture adjustments suitable for different situations. In addition, we considered natural gaze motion, such as the convergence of both eyes and the ratio of eye angle to head angle, based on the knowledge of humans and other social animals (Kondo, Kawamura, Takemura, Takamatsu, & Ogasawara, 2011).

In this paper, we develop a HRI system to embed gaze and gestures. The system can communicate smoothly with multiple people by rapidly switching from person to person and adjusting a gesture to the location of a speaker and/or object as shown in Figure 1. These functions are also applied even with one-to-one HRI, since we assumed one-to-one HRI is the subset of multi-party HRI. The number of situations that our system can deal with is exponentially increased due to combinatorial explosion as compared to conventional one-to-one HRI systems. We solved this problem by cooperation between a Key-Value Store, which has ACID (Atomicity, Consistency, Isolation, and Durability) properties and other system components. Therefore, we do not have to be concerned about the synchronization and scalability of the system.

The rest of this paper is organized as follows: Section 2 describes related works. Section 3 describes how to plan body gestures and facial expressions for an android. The details of the rest of the HRI system components are described in Section 4. To evaluate the effectiveness of our system, we conducted HRI experiments on a total of 1,662 visitors. Section 5 shows the results of the speaker ratio (i.e., the ratio of the number of people who start speaking to the robot, to the number of people who approach the robot) and the residence time of their communication. Section 6 shows the results of the analysis of human impressions of the android. Finally, we conclude the paper and describe future work in Section 7.
2. Related Works

Previous research methods in HRI mainly focus on one-to-one communication. Researchers Ido et al. (Ido, Matsumoto, Ogasawara, & Nishimura, 2006) developed a Q&A communication system that reacts based on a human’s speech and gaze tracking. Chikaraishi et al. (Chikaraishi, Minato, & Ishiguro, 2008) proposed to add an ability for a natural idling motion (i.e., a robot moves based on an Attractor Selection Model (Kashiwagi, Urabe, Kaneko, & Yomo, 2006)) with online human tracking. Sakamoto et al. (Sakamoto et al., 2007) developed a telecommunication system that can remotely control an android using facial motion capture. These research methods were effective in a limited task domain; however, the expansion of the task domain makes the preparation of all the gestures unfeasible.

For multi-party communication, researchers have speculated about how a robot can participate in multi-party conversations (Matsusaka et al., 2003; Matsusaka, S. Fujie, & Kobayashi, 2010). This system could estimate a current speaker and the next speaker in the party by gaze tracking and speech recognition, reacting to improve awareness of the robot. Nakanishi et al. (Nakanishi, Nakazawa, Ishida, Takenashi, & Isbister, 2003), and Sakamoto and Ono (Sakamoto & Ono, 2006) proposed how to construct relations between agents and humans or between robots and humans using the psychology of interpersonal relations. These research methods are mainly focused on recognizing the multi-party conversational situation. We are now developing a system on which the RMDB is implemented to enable not only recognition, but also interaction for multiple people with human-like gestures and facial expressions.

3. Planning Body Gesture and Facial Expression

3.1 Actroid-SIT

In HRI, we consider human impressions affected by the appearance of a robot as well as by the robot’s body gestures. In this research, we use an android Actroid-SIT made by Kokoro Company, Ltd. The Actroid has a human-like figure (as shown in Figure 2) and total of 42 degree of freedoms (DOFs) (29 DOFs for body gesture and 13 DOFs for facial expression, as shown in Table 1). Since the Actroid is driven by pneumatic actuators, damage caused by an accidental collision is much less than with other motor-driven rigid robots.
3.2 Planning Body Gesture

Our proposed motion planning method, a *Reconfigurable Motion Database* (RMDB), can generate flexible gestures in real-time. The RMDB is based on a *Motion Graph* (Kovar, Gleicher, & Pighin, 2002) and a *Match Web* (Kovar & Gleicher, 2004) which are proposed by Kovar et al. Originally, the *Motion Graph* and *Match Web* were aimed at CG animation. To apply these methods to a real robot, and not to a CG avatar, however, we needed to add the ability of real-time collision avoidance. The RMDB has the following two main features: motion interruptivity and motion parameterization.

3.2.1 Motion Interruptivity

In our previous study (Kondo et al., 2010), we surveyed that speaker-intended interruption occurs in 26% of communications for some reasons, such as a switch of speaker’s interest, a nod in response, speaker replacement, and speech recognition failure. Whenever a speaker interrupts a robot’s current reaction, it is necessary to suddenly terminate the social task and then transition appropriately to the next reaction. Given two motion sequences \( A \) and \( B \), the algorithm for the motion transition first calculates the similarity between poses \( a_i \) and \( b_j \), in sequences \( A \) and \( B \). When a Euclidean distance \( D(a_i, b_j) \) is locally minimized, the algorithm attempts to connect two motions around these two poses, where \( i, j, \) and \( k \) are frame indexes, and \( m \) is a user-defined number of frames for connection, as follows:

\[
c_k = \alpha(k) a_{i+k} + (1 - \alpha(k)) b_{j-m+k}, \\
\alpha(k) = 2 \left( \frac{k+1}{m} \right)^3 - 3 \left( \frac{k+1}{m} \right)^2 + 1.
\]

If the Euclidean distance \( D(a_i, b_j) \) is larger than a threshold for all \( i \) and \( j \), then two motions are connected by an interpolation path through online planning (Kondo et al., 2010).

This algorithm is not only for interruptible gestures, but also utilized for a series of motion primitives such as a single connected gesture.

3.2.2 Motion Parameterization

Given a human-like motion sequence that is easily obtained from the motion capture system, the RMDB can extract similar sequence regions from all of the sequences by the function of the *Match Web* (Kovar & Gleicher, 2004). Then, given a target parameter \( \tilde{p} \) (e.g., the 3D position of hand or the angle of face) against the gestures, the target motion is obtained by adding all the similar motions with a given weight. The weight \( w_i \) of the \( i \)-th similar motion is given by equation (3), whereby \( c \) is a constant related to variance of \( \tilde{p} \), and \( k \) is the number of similar motions for each gesture shown in the third column of Table 2.

\[
w_i = \frac{\exp \left( -\frac{D(\tilde{p}, p_i)^2}{c} \right)}{\sum_{j=1}^{k} \exp \left( -\frac{D(\tilde{p}, p_j)^2}{c} \right)}.
\]

In this paper, we measured daily interaction motions of five male and female subjects who were in their 20s. A total of 15-minutes motions were used for the construction of the RMDB. The sensor suit MVN made by Xsens Technologies, Inc. was used for measurement of motions at 30 [Hz]. The RMDB has 18 gestures, including more than three similar motions in the 15-minutes motions, as shown in Table 2. The terms \( x, y, z \) indicate the 3D position of a target. The figure \( \phi \) indicates the yaw (horizontal) angles while \( \theta \) indicates the pitch (vertical) angles of the Actroid’s direction, and \( l \) indicates the breadth between both hands. Gestures #9, #10, #11, #12, #16, #17, #18 use only one hand and are archived by either the left hand or the right hand. The RMDB can select an appropriate hand based on the target parameter \( \tilde{p} \) and gesture #1 performs to track a speaker’s direction by combining gaze, as described in Section 3.3.2.
Table 2: Parameters and the number of similar motions for each gesture. The terms $x, y, z$ in parameter $\tilde{p}$ indicate the 3D position of a target. The figure $\phi$ indicates the yaw (horizontal) angles while $\theta$ indicates the pitch (vertical) angles of the Actroid’s direction, and $l$ indicates the breadth between both hands.

<table>
<thead>
<tr>
<th>gesture</th>
<th>parameter $\tilde{p}$</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 GAZING</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>#2 YES / NODDING</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>#3 NO</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>#4 BOWING</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>#5 BOTH_BYEBYE</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>#6 SHAKING_NeCK</td>
<td>($\phi, \theta$)</td>
<td>4</td>
</tr>
<tr>
<td>#7 THINKING</td>
<td>($\phi$)</td>
<td>6</td>
</tr>
<tr>
<td>#8 FLEXING_MUSCLE</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>#9 POINTING_AT_MYSELF</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>#10 BYEBYE</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>#11 SWINGING</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>#12 LOOKING_AT_CLOCK</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>#13 SHAKING</td>
<td>($\phi, \theta, l$)</td>
<td>12</td>
</tr>
<tr>
<td>#14 BOTH_SHAKING</td>
<td>($\phi, \theta, l$)</td>
<td>11</td>
</tr>
<tr>
<td>#15 BOTH_SPREADING</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>#16 POINTING_UP_DOWN</td>
<td>($x, y, z$)</td>
<td>10</td>
</tr>
<tr>
<td>#17 PUNCHING</td>
<td>($x, y, z$)</td>
<td>10</td>
</tr>
<tr>
<td>#18 POINTING</td>
<td>($x, y, z, \phi, \theta$)</td>
<td>26</td>
</tr>
</tbody>
</table>

Figure 3 shows examples of pointing gestures (#18) where the target parameter $(x, y, z)$ is shifted from the right side to the left side horizontally. This figure shows precise motion synthesis is achieved while keeping the twisting motions of the wrist, which simulates actual human motion. We have already discussed the accuracy of the synthetic motions in our previous study (Kondo, Takemura, Takamatsu, & Ogasawara, 2012).

3.3 Planning Facial Expression

The Actroid can control facial expression and gaze direction in the same way as a human. Retargeting the human’s facial expression into the Actroid’s facial expression is very difficult because of the complexity of the actuator. Therefore, in our system, the module to control facial expression and gaze is implemented independently with the RMDB. In addition, speech generation is also a part of this planning.

3.3.1 Facial and Speech Expression Since the Actroid has the eyebrow, eyelid, and cheek joints for controlling facial expression, in this paper we used the Actroid to perform simple facial expressions (as shown in Figure 4). The Actroid has five emotions: NORMAL, SAD/ANGRY, and JOYFUL/HAPPY. To distinguish between SAD/ANGRY, and JOYFUL/HAPPY, we also changed voice parameters ($i.e.$, speed, pitch, volume, and pause time) to make the expressions unique using a voice synthesis software called VoiceText¹, made by HOYA Service Corporation.

¹http://voicetext.jp/
Figure 3. Synthetic pointing gestures based on pointing horizontal location $x$ in the parameter $(x, y, z, \phi, \theta)$ where $-200 \, [\text{mm}] \leq x \leq 300 \, [\text{mm}]$. Since the Actroid’s horizontal center is $x = 0 \, [\text{mm}]$ and its shoulder-width is 450 [mm], the pointing locations were executed with satisfactory accuracy.

Figure 4. Comparison of facial expressions: (a) SAD/ANGRY, (b) NORMAL, and (c) JOYFUL/HAPPY. The Actroid can control only the eyebrow, eyelid, and cheek for facial expressions.
Figure 5. The definition of the angles of the head and both eyes is done by considering the sharing rate and convergence.

\[
V_x = -\frac{\pi}{3}
\]

\[
V_x = -\frac{\pi}{6}
\]

\[
V_x = +\frac{\pi}{6}
\]

\[
V_x = +\frac{\pi}{3}
\]

Figure 6. Examples of the facing angle \(V_x\) = (a) \(-\frac{\pi}{3}\), (b) \(-\frac{\pi}{6}\), (c) \(+\frac{\pi}{6}\), (d) \(+\frac{\pi}{3}\), respectively

### 3.3.2 Gaze Motion

Gaze angle \(V_x\) indicates the angle between the front direction and the gaze direction as shown in Figure 5. \(V_x\) is the sum of the head angle \(H_x\) between the head direction and the gaze direction \(E_x\). The sharing rate \(D_x = \frac{E_x}{V_x}\) indicates the ratio of eye movement to head movement in gaze motion. Although the sharing rate is altered by ages and/or genders, in this paper, we employed the sharing rate as defined in equation (4), which Masuko et al. (Masuko & Hoshino, 2005) designed for a CG avatar.

\[
D_x = \begin{cases} 
1 & |V_x| \leq \frac{\pi}{12} \\
\frac{|E_x|}{|V_x|} & \frac{\pi}{12} < |V_x| \leq \frac{\pi}{6} \\
0 & otherwise 
\end{cases}
\]  

(4)

Eye convergence occurs due to a gap in the direction from both eyes, when gazing at an object located a finite distance. Let \((x, y, z)\) be an object position. Given that the direction of the object is \(E_x\), the direction of left eye is \(E_{xl}\), and that of the right eye is \(E_{xr}\), then \(E_{xl} + E_{xr} = E_x\) can be satisfied. We can then generate angles \(V_x\) and \(H_x\) considering the sharing rate and convergence with the following equations:
\[ V_x = \tan^{-1} \left( \frac{x}{z} \right), \tag{5} \]
\[ H_x = (1 - \alpha D_x) V_x, \tag{6} \]

where \( \alpha (0 \leq \alpha \leq 1) \) is for adjustment of the convergence ratio, which is empirically assigned.

We used \( \alpha = 0.5 \), which is the best precision for perception of the gaze direction in our preliminary experiments (Kondo et al., 2011). Next, given an object point \((\hat{x}, \hat{y}, \hat{z})\) calculated by \((x, y, z)\) rotating \(H_x\) along head coordinates, we measured \((\hat{x} - \frac{1}{2}, \hat{y}, \hat{z} + d)\) as the length from the right eye to the object and \((\hat{x} + \frac{1}{2}, \hat{y}, \hat{z} + d)\) as the same for the left eye. We calculated angles \(E_{xl}\) and \(E_{xr}\) by equations. (7) and (8). Finally, we planned the gaze motion as shown in Figure 6. There are examples of angle \(V_x\) as \(-\frac{\pi}{3}, -\frac{\pi}{6}, +\frac{\pi}{6}, +\frac{\pi}{3}\), respectively.

\[ E_{xl} = \tan^{-1} \left( \frac{\hat{x} + \frac{1}{2}}{\frac{\hat{z}}{2}} \right), \tag{7} \]
\[ E_{xr} = \tan^{-1} \left( \frac{\hat{x} - \frac{1}{2}}{\frac{\hat{z}}{2}} \right). \tag{8} \]

4. Implementation

Figure 7 shows the system configuration of our proposed system. This system has six components: a Key-Value Store, Episode Rule Selector, Speaker Detector, Speech Recognizer, Motion Planner, and a Dialogue Generator. Note that the Motion Planner and the Dialogue Generator have already been described in Section 3.

4.1 Key-Value Store

The Key-Value Store is one of the dictionary data structures that have a set of pairs, which include an index (Key) and the related value (Value). The Key-Value Store has been developed as an alternative to the Relational Database (Codd, 1970) which is inferior due to an overhead of network communication and the difficulty of parallel processing. Since the Key-Value Store uses a distributed hash table, it is easy to process in parallel with redundancy. We used a Redis \(^2\), which is one of the Key-Value Store implementations, and let it perform as a web server. All components in the HRI system access this Key-Value Store using a RESTful API (Fielding, 2000). This API is thread-safe and allows dynamic changes of connection among modules.

Various architectures have been proposed for robot systems, and the stability and flexibility have been studied in these architectures. In recent architectures, OpenRTM (Ando, Suehiro, Kitagaki, Kotoku, & Yoon, 2005) and ROS (Quigley et al., 2009) have become popular. Since they provide operating system-like functionality, they are heavy and inconvenient to use only for connections and communication among the components of the system. Therefore, we applied the Key-Value Store to the robot system as a centralized administrative framework.

4.2 Episode Rule Selector

The Episode Rule Selector decides the reaction of the Actroid based on an episode rule database when sensory data changes. Kanda et al. (Kanda, Shiomi, Miyashita, Ishiguro, & Hagita, 2009)

\(^2\)http://redis.io/
originally proposed the idea of the Episode Rule Selector. In this paper, the episode rules are described using a script language, Jython. Therefore, the rules can be modified and added online even while the system is running. Each episode rule contains the following three functions:

- **Precondition**
  Returns a score calculated based on the speaker’s dialogue, existence of the speaker, and the number of people, etc., which are obtained from the Key-Value Store.

- **Interaction**
  Configures a gesture’s type and its parameter, and sentences in reply, and emotion. In the case of Figure 7, the Actroid replies to, “What’s your name?” with two gestures to promote the next interaction.

- **Aftereffect**
  Terminates this rule, and sets or resets several condition flags in the Key-Value Store for the next interaction. The key “already asked name” can be seen in Figure 7. This key is the flag for the next continuous interaction “My name is ...”, which we expected.

The algorithm of Episode Rule Selector is shown in Figure 8. While the system is running, the algorithm repeats the following procedures. First, the episode rule database $\mathcal{R}$ is loaded. Then, the system selects the episode rule $r$ which has a maximum score of the precondition in $\mathcal{R}$, and the functions of $r$.interaction and $r$.aftereffect are called in order. This interaction loop can be interrupted even if a previous $r$.interaction is not finished. Current gesture can be transitioned to the next gesture smoothly thanks to motion interruptivity, and the VoiceText can interrupt current voice while resuming with the next voice. We can accelerate reaction speed by planning the interruption-aware interaction.

---

Figure 7. System configuration of the multi-party HRI system with six components: a Speaker Detector and a Speech Recognizer are for sensing; a Dialogue Generator and a Motion Planner are for the Actroid, and an Episode Rule Selector is for planning how the Actroid should interact. All components communicate via a Key-Value Store.
while system is running do
\[ \mathcal{R} \leftarrow \{ \mathcal{R} \cup \text{updated or added rules} \} \]
\[ r = \arg \max_{r_i \in \mathcal{R}} \text{precondition}. \]
\[ r.\text{interaction} \]
\[ r.\text{aftereffect} \]
end while

Figure 8. The algorithm of the Episode Rule Selector implemented on the proposed HRI system. The terms \( \mathcal{R} \) indicates the episode rule database and \( r \) indicates the system-selected rule.

Figure 9. An example of the image processing, where the circles indicate the results of face detection and the square indicates the result of color extraction for microphone detection.

Our system can deal with action episode rules showing that the robot actively performs an interaction when there is no speaker, as well as reaction episode rules that are passive rules when the speaker asks a question.

4.3 Speaker Detector

To detect the speaker’s position and to count the number of people around the Actroid, we use an IEEE 1394 camera (Sony DFW-VL500) located next to the Actroid. Figure 9 shows the result of the image processing. Circles in this figure indicate the result of face detection by Haar-like features (Viola & Jones, 2001) in the OpenCV 4 library. A square also indicates the result of a color extraction of tapes attached to the microphone. It is difficult to detect a speaker (i.e., the person who is speaking) only with only facial information. Therefore, we assumed a person who has a microphone is the speaker. We used two different colors (yellow and blue) for the detection.

4.4 Speech Recognizer

We employed Julius 5 as speech recognition software and the Yahoo Key-phrase Extraction Web API 6 to extract important keywords (which are a subset of the sentence recognized by Julius). The keyword sometimes consists of multiple words. The API scores each keyword based on its importance. In Figure 7, the sentence, “Who are you?” is divided into the keywords “you,” “who,” and the sentence, “Who are you?” For robust matching, the whole sentence is added as a low priority keyword.

In Japanese especially, it is very difficult to grammatically recognize the speaker’s speech. However, this system employs a word spotting approach with scored keywords. For example, in case of

---

5http://sourceforge.net/projects/opencvlibrary/
6http://julius.sourceforge.jp/
6http://developer.yahoo.co.jp/webapi/jlp/keyphrase/v1/extract.html
Table 3: Sentences and the sets of their keywords and scores for replying to “Who are you?” Higher score indicates the keyword is more important. 100 is the highest score and 0 is the lowest score.

<table>
<thead>
<tr>
<th>sentence</th>
<th>the set of keyword and score</th>
</tr>
</thead>
<tbody>
<tr>
<td>My name is Kondo.</td>
<td>{“Kondo”=100, “name”=46, “my”=21}</td>
</tr>
<tr>
<td>Please call me Kondo.</td>
<td>{“Kondo”=100, “call”=39, “me”=21}</td>
</tr>
<tr>
<td>I am Kondo.</td>
<td>{“Kondo”=100, “I”=21}</td>
</tr>
<tr>
<td>This is Kondo.</td>
<td>{“Kondo”=100, “this”=18}</td>
</tr>
<tr>
<td>Kondo.</td>
<td>{“Kondo”=100}</td>
</tr>
</tbody>
</table>

relying to: “What is your name?” Table 3 shows four answers that mean almost the same (“My name is Kondo,”) but their grammar and the way of speaking are different. However, their primary keywords are the speaker’s name “Kondo” in all cases, so the Actroid can reply: “Your name is Kondo, isn’t it?”.

5. Evaluation of Speaker Ratio and Residence Time

We evaluated the effectiveness of the proposed method, i.e., motion interruptivity and motion parameterization. To do that, we compared the proposed system to the systems where motion interruptivity and/or motion parameterization were not implemented.

5.1 Hypotheses and Predictions

The proposed motion interruptivity and parameterization make the HRI system more responsive and active. We believe people feel comfortable to communicate through the system, because the motion parameterization makes the Actroid active, thanks to gestures which have $\phi$, $\theta$ parameters, such as idling motion by gesture #1 GAZING. This is why the Actroid is busy looking around at people, and subjects might feel easy talking to the Actroid. In addition, the motion interruptivity makes for responsive communication (Kondo et al., 2010). Thus, the response time (i.e., the latency of communication) is decreased; nevertheless the communication can be lively and durable. We expected that a better HRI system gets a higher speaker ratio and longer residence time. Therefore, we made the following predictions:

- **Prediction 1**: Motion parameterization will increase the number of people who voluntarily speak to the Actroid.
- **Prediction 2**: Motion interruptivity will increase the residence time of the communication with the Actroid.

5.2 Conditions

We controlled the following two conditions: the motion interruptivity (interruptible) and the motion parameterization (parametric). Combining these conditions, there are four comparative systems, UI+NP, I+NP, UI+P, and I+P. The systems are shown in Table 4. Each system is indicated as follows:

- **UI+NP**: a conventional interaction system that does not consider both motion interruptivity and parameterization.
- **I+NP**: the one-to-one interaction system that we developed in previous research (Kondo et al., 2010).
- **UI+P**: can parameterize gestures but cannot interrupt interaction.
- **I+P**: the multi-party interaction system, which we proposed in this paper.
Of course the systems UI+NP, I+NP, and UI+P can interact with multiple people, but UI+NP and UI+P cannot interrupt the interaction, and UI+NP and I+NP can interact with people assuming that they are always in front of the Actroid.

5.3 Method

5.3.1 Participants Experiments were conducted for four weekdays in the Heijo palace site, Japan. One of the above-mentioned systems was evaluated on one of the weekdays. 1,662 visitors in total were took part in these experiments as the subjects. Table 4 shows the number of the speakers, non-speakers, and their totals (i.e., subjects) for each day. The subject in the table is the number of people who approach the Actroid, the speaker is the number of subjects who voluntarily spoke to the Actroid, and the non-speaker is the number of subjects who did not speak to the Actroid (i.e., subject = speaker + non-speaker). Subjects were in various age groups and had a similar age distribution on each day.

5.3.2 Settings We used the Actroid as a receptionist and used question and answer (Q & A) communication, such as Q: “Where is the bathroom?” A: “It is on your right.” The Actroid can also communicate dozens of daily conversations such as Q: “How old are you?” A: “I was born three years ago. And you?” Figure 10 shows a snapshot of the communication when a speaker asked a question to the Actroid.

We assumed that among the subjects, only one subject would be allowed to speak to the Actroid

Table 4: The number of subjects, speakers and non-speakers with respect to four comparative systems. The terms interruptible/un-interruptible means the motion interruptivity is enabled/disabled and parametric/non-parametric means the motion parameterization is enabled/disabled.

<table>
<thead>
<tr>
<th>system</th>
<th>condition</th>
<th>speaker</th>
<th>non-speaker</th>
<th>subjects (total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UI+NP</td>
<td>un-interruptible and non-parametric</td>
<td>226</td>
<td>241</td>
<td>467</td>
</tr>
<tr>
<td>I+NP</td>
<td>interruptible and non-parametric</td>
<td>165</td>
<td>267</td>
<td>432</td>
</tr>
<tr>
<td>UI+P</td>
<td>un-interruptible and parametric</td>
<td>263</td>
<td>166</td>
<td>429</td>
</tr>
<tr>
<td>I+P</td>
<td>interruptible and parametric</td>
<td>206</td>
<td>128</td>
<td>334</td>
</tr>
</tbody>
</table>
5.4 Measurement

Two measurements were conducted:

- **Speaker ratio** is defined as the ratio of subjects who spoke to the Actroid (i.e., $\frac{\text{speaker}}{\text{subject}}$).
- **Residence time** is defined as the duration from the moment when a subject starts speaking to the moment when the subject puts his/her microphone back onto its stand.

5.5 Results

First, we verified **Prediction 1**. Figure 11 shows the results of speaker and non-speaker for each system. A chi-square test was revealed significant differences among conditions ($\chi^2(3) = 62.765, p < .01, \phi = 0.194$). A residual analysis revealed the following: speaker in UI+NP is significantly low (residual $= -1.79, p < .10$) and non-speaker in UI+NP is significantly high (residual $= 1.79, p <$ \( \footnote{We assumed that in many cases, putting the microphone back represented losing interest in the Actroid.} \) at a time. This assumption is plausible because it is difficult, even for humans, to answer questions asked by multiple people simultaneously. We regarded a single person as a special case of multi-party HRI; we did not control the number of attendees at one time.

Figure 11. Comparison of speaker and non-speaker between four experimental systems.

Figure 12. Comparison of the mean and the standard error of the residence time between four experimental systems.
Figure 13. Snapshots of HRI experiments with the system I+P where a speaker is located at the (a) right side, (b) center, and (c) left side of chairs.

speaker in I+NP is significantly low (residual = -6.552, p < .01) and non-speaker is significantly high (residual = 6.552, p < .01) in I+NP. speaker in UI+P is significantly high (residual = 4.064, p < .01) and non-speaker is significantly low (residual = -4.064, p < .01) in UI+P. speaker in I+P are significantly high (residual = 4.601, p < .01) and non-speaker in I+P are significantly low (residual = -4.601, p < .01). These results implicitly indicates that motion parameterization increases speaker ratio.

Next, we verified Prediction 2. Figure 12 shows the mean and the standard error of the residence time for each system. A 2 × 2 two-way repeated-measure analysis of variance (ANOVA) was conducted (N = 208.96). A significant main effect in interruptible was revealed (F(1, 856) = 39.22, p < .01, η² = .044). Thus, we confirmed that the motion interruptivity makes HRI a more durable form of communication.

6. Evaluation of Human Impression

In previous Section 5, we confirmed the motion interruptivity makes communication more durable while motion parameterization makes people more easily approach the Actroid. In this section, we evaluate the effectiveness of the motion parameterization after approaching the Actroid (i.e., during communication).

6.1 Hypothesis and Prediction

The proposed motion parameterization makes people more easily approach the Actroid. We believe this result is caused by people perceiving a more positive impression of the Actroid. Thus, we made the following prediction:

• Prediction: Due to the motion parameterization people will perceive more positive impressions, that is, a better impression, toward the Actroid.

6.2 Conditions

In this section, we controlled only one condition, the motion parameterization. The interruptivity was always enabled, because we have already confirmed the interruptivity makes more responsive and durable communication. We observed the result to mean that people perceived a better impression of the Actroid. Thus, we measured human impressions of the Actroid with two systems—I+NP and I+P—as mentioned in Section 5.2.

6.3 Method

6.3.1 Participants Experiments were conducted with 42 subjects of various age groups (20 subjects for the I+NP condition and 22 subjects for the I+P condition). All subjects were visitors who attended an open house day of our university and voluntarily answered a questionnaire at the end of
the communication with the Actroid. One of the systems was changed to another system every one hour. We let subjects interact freely, without suggesting anything (such as sitting position or timing during the communication process).

6.3.2 Settings Subjects sat down on one of three seats placed in front of the Actroid and then communicated with each other for a few minutes. Figure 13 shows the snapshots of the communications with the condition I+P, when the speaker sat on the (a) right-side seat, (b) center seat, and (c) left-side seat. As shown in Figure 13, the head and body of the Actroid faced the speaker’s direction by motion parameterization. The other settings are the same as described in Section 5.3.2.

6.4 Measurement We used the Semantic Differential (SD) method (Osgood & Tannenbaum, 1957) for the evaluation of human impressions of the Actroid. After the communication with the Actroid was finished, each subject answered a questionnaire with 28 antonymous adjective pairs on a Likert scale from one to seven points (i.e., one: the worst, seven: the best), for an SD profile. Each adjective pair is shown in Table 5. The higher score represents a positive or better impression.

6.5 Results The results of the mean and the standard error of the SD profiles are shown in Figure 14. This figure shows that almost all scores in I+P condition are higher. Through detailed analysis we found that the parameterization increases not only the speaker’s score, but also his or her neighbor’s score.

Next, a factor analysis was conducted (eigenvalue $\geq 1$, cumulative variance $\leq 50\%$, factor loading $\geq 0.5$). Figure 15 shows four factor scores by Bartlett method (Bartlett, 1937) and Promax rotation (Hendrickson & White, 1964). Each factor contains some adjectives as shown in Table 6. Note that we named each factor based on its adjectives.

A student t-test revealed that activity ($p < .01$), sophistication ($p < .01$), speediness ($p < .01$), and friendliness ($p < .1$) scores in I+P condition were significantly higher than ones in I+NP condition. Therefore, our hypothesis that the motion parameterization makes more positive impression

Table 5: 28 antonymous adjective pairs (left side: positive, right side: negative) described in the questionnaire

<table>
<thead>
<tr>
<th>positive</th>
<th>negative</th>
<th>positive</th>
<th>negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>good</td>
<td>bad</td>
<td>sensitive</td>
<td>insensitive</td>
</tr>
<tr>
<td>kind</td>
<td>afraid</td>
<td>fulfilling</td>
<td>empty</td>
</tr>
<tr>
<td>cute</td>
<td>hateful</td>
<td>bright</td>
<td>dark</td>
</tr>
<tr>
<td>fun</td>
<td>boring</td>
<td>active</td>
<td>passive</td>
</tr>
<tr>
<td>warm</td>
<td>cold</td>
<td>familiar</td>
<td>unfamiliar</td>
</tr>
<tr>
<td>approachable</td>
<td>unapproachable</td>
<td>fast</td>
<td>slow</td>
</tr>
<tr>
<td>humanly</td>
<td>mechanically</td>
<td>quick</td>
<td>dull</td>
</tr>
<tr>
<td>cheerful</td>
<td>awful</td>
<td>interesting</td>
<td>uninteresting</td>
</tr>
<tr>
<td>friendly</td>
<td>unfriendly</td>
<td>considerate</td>
<td>selfish</td>
</tr>
<tr>
<td>likable</td>
<td>dislikable</td>
<td>complicated</td>
<td>simple</td>
</tr>
<tr>
<td>positive</td>
<td>negative</td>
<td>safe</td>
<td>dangerous</td>
</tr>
<tr>
<td>affable</td>
<td>disgusting</td>
<td>comprehensive</td>
<td>incomprehensive</td>
</tr>
<tr>
<td>wise</td>
<td>stupid</td>
<td>intense</td>
<td>mild</td>
</tr>
<tr>
<td>flashy</td>
<td>plain</td>
<td>strong</td>
<td>weak</td>
</tr>
</tbody>
</table>
Figure 14. The mean and the standard error of the SD profiles of $I+NP$ and $I+P$ scores. We used a Likert scale from one to seven points (i.e., one: the worst, seven: the best) for the questionnaire.

Figure 15. The mean scores and the standard errors of the four factors: friendliness, activity, sophistication, and speediness were obtained by analyzing the SD profiles shown in Figure 14. The higher score represents a more positive impression.
Table 6: Four factors given by the factor analysis and their containing adjectives (positive ones only)

<table>
<thead>
<tr>
<th>factor</th>
<th>adjectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>friendliness</td>
<td>cute, like, friendly, cheerful, positive, kind, fun</td>
</tr>
<tr>
<td>activity</td>
<td>active, fulfilling, sensitive, considerate</td>
</tr>
<tr>
<td>sophistication</td>
<td>comprehensive, approachable, wise</td>
</tr>
<tr>
<td>speediness</td>
<td>flashy, quick, fast, warm</td>
</tr>
</tbody>
</table>

was supported. This result might indicate that the subjects implicitly perceived the Actroid as being wiser and moving more quickly due to motion parameterization. Therefore, we concluded that body gesture planning is one of the most useful functions for more human-like HRI systems.

7. Conclusions

In our research, we focused on online body gesture planning for androids based on a Reconfigurable Motion Database (RMDB) in HRI. Communication with multiple people is more common than one-to-one communication in the real world. Our proposed method RMDB has motion interruptivity which can transition from gesture to gesture smoothly and react rapidly, as well as having motion parameterization, which can adjust motion suitable for human demand. Thanks to this advantage, an android can communicate with multiple people by rapidly switching to each person, and adjust a gesture to the location of a speaker and/or object.

We developed a novel interaction system on an android Actroid-SIT which is embedded into the RMDB for body gesture. In addition, our system can generate facial expression with five emotions, including gaze motion based on the knowledge of social animals. The system architecture is inspired by an Episode Rule Selector (Kanda et al., 2009), and its components are connected via a Key-Value Store. The Key-Value Store has ACID properties. Therefore, we solved the issue of synchronization among components.

To compare a speaker ratio (i.e., the ratio of the number of people who start speaking to the Actroid, to the number of people who approach the Actroid) and the residence time of communication with or without the motion interruptivity and parameterization that we proposed, we conducted multi-party HRI experiments for 1,662 subjects in total. With our HRI system, the speaker ratio was over 60%, while that of conventional systems was less than 50%. The residence time of communication was longer in our HRI system. Thus, interruptivity makes for more durable HRI communication, and as well as more responsive HRI communication.

By analyzing human impressions of the Actroid, we proved that motion parameterization contributed to the Actroid being wiser and more comprehensive. In these experiments, the Actroid generated appropriate gestures using the parameterization of the RMDB and faced in the speaker’s direction. As a result of the SD method and factor analysis, a 1% level of significant differences between parametric and non-parametric gestures existed in activity, sophistication, and speediness factors. We also found that the way of communication is dissimilar among age, gender, and character; future work will evaluate these dissimilarities.

Furthermore, we need to make the system more advanced by adding and/or upgrading its components. We must also consider the automatic creation of the episode rules by through a human history of tele-operation based on a Wizard of Oz method (Fraser & Gilbert, 1991).
Acknowledgments

This work was supported by Japan Grant-in-Aid for Scientific Research on Innovative Areas No. 22118507.

References


Authors’ names and contact information: Yutaka Kondo, Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan. Email: yutaka-ko@is.naist.jp. Kentaro Takemura, Faculty of Information Science, Nara Institute of Science and Technology, Ikoma, Japan. Email: kenta-ta@is.naist.jp. Jun Takamatsu, Faculty of Information Science, Nara Institute of Science and Technology, Ikoma, Japan. Email: j-taka@is.naist.jp. Tsukasa Ogasawara, Faculty of Information Science, Nara Institute of Science and Technology, Ikoma, Japan. Email: ogasawar@is.naist.jp.