Studies on grounding with gaze and pointing gestures in human-robot-interaction

Markus Häring, Jessica Eichberg, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Häring, Markus, Jessica Eichberg, and Elisabeth André. 2012. "Studies on grounding with gaze and pointing gestures in human-robot-interaction." In *Social robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012*, edited by Shuzhi Sam Ge, Oussama Khatib, John-John Cabibihan, Reid Simmons, and Mary-Anne Williams, 378–87. Berlin [u.a.]: Springer. https://doi.org/10.1007/978-3-642-34103-8_38.

Studies on Grounding with Gaze and Pointing Gestures in Human-Robot-Interaction

Markus Häring, Jessica Eichberg, and Elisabeth André

Institute of Computer Science, Human Centered Multimedia, Augsburg University, Germany

Abstract. In this study we investigated the use of gaze and pointing gestures in scenarios where a human has to follow the instructions of a humanoid robot. Our objective was to analyze the performance of a human participant, that solves an abstract jigsaw puzzle with the help of our robot instructor, in different grounding scenarios with varying difficulty. Furthermore we investigated how the attitude towards the robot and the self-assessment of the participant changed. Our results support that adding gaze to the interaction usually improves the interaction, but often additional pointing gestures are needed to make a significant difference.

1 Introduction

Social robots are - among other duties - supposed to make our lives easier by assisting us with more or less complex tasks. When this assistance is based on collaboration the robot might also have to take the role of an instructor. For example, people who are not at all versed in a manual skill could be showed by a robot how to assemble a wardrobe. A robot would give them step by step instructions, which part of the wardrobe has to be installed with which tool at what position. A complex interaction like this involves a lot of coordination not only of physical tasks but also of conversational actions. The process of updating mutual knowledge, mutual assumptions and mutual beliefs during the interaction is called *grounding* [3]. To minimize the chance of errors during this process due to misunderstandings humans extend their verbal utterances with gaze and pointing gestures and social robots will have to rely on that modalities as well.

Sugiyama and Kanda [8] already confirmed that deictic gestures help robots correct misunderstood verbal utterances. Faber and Bennewitz [4] stated that, in a conversation, a robot should keep eye-contact with its human interlocutor in order to show him attention. Moreover a robot can use gaze to target and clarify the object of interest a conversation is dealing with. But even though many applications for social robots combine gaze and pointing gestures, no detailed comparison investigating the benefits of the two modalities has been done so far. Under which conditions is gaze behavior additional to speech sufficient for stable grounding and when is it mandatory to include pointing gestures to avoid

critical errors? Will the users even pay attention to these cues while they are concentrating on the task? How does the user's attitude towards the robot change with changing level of modality also considering the complexity of the task?

In the presented study the participants were exposed to a situation they were not able to solve on their own, consequently they had to rely on the guidance of a robot. As a generalized testing environment we created an abstract puzzle game for the touch sensitive Microsoft Surface. In this game the robot acted as an *instructor* and guided the human participant through the whole puzzle job using either only speech, speech combined with gaze or speech combined with gaze and pointing gestures. We investigated the robot's performance as an instructor by logging the number of mistakes made by the participant and the time needed to solve the puzzle. We also asked the participants about their attitude towards the robot after their interaction and were also interested in their self-assessment regarding their own performance during the puzzle game. With the last aspect we wanted to evaluate to which extent the used modalities affect the participants self-efficacy when solving the puzzle.

As the use of additional modalities is also a question of efficiency (regarding energy consumption, hardware requirements and implementation workload) this study wants to clarify if and when this effort is necessary and what the benefits are, also in regard to improvements of the user's performance.

2 Related Work

A number of investigations dealing with human-robot interaction and puzzle games can be found in the literature that are of relevance for our study.

Burghart and Gaertner [2] analyzed the cooperative solving of a jigsaw puzzle between robot and human tutor. In contrast to our study, where the robot is taking on the role of an instructor, it is the human who provides support to the robot in their study, but only when the human tutor evaluates the last action of the robot as negative.

Giuliani and Knoll [5] observed how participants interacted in an cooperative construction task with an instructive robot, compared to the interaction with a supportive robot. While the instructive robot first instructs the user how to proceed with the construction and then supports the user by handing over building pieces, the supportive robot keeps a more passive role and only intervenes when the user is about to make a mistake. The subjective and objective data of their evaluation suggests that participants don't prefer one of the different roles rather will they adapt to the situation by taking the counterpart to the robot's role. Giuliani and Knoll's study focuses on the evaluation of the different roles and does not investigate the effects of the used modalities in their robot behaviors.

Salem et al. [7] considered multiple modalities in their study and installed a robot in a household scenario, where it assisted a human participant by providing information. The participants had to place some kitchen items in a cupboard while they had to pay attention to the robot's instructions. The following two conditions according to the robot's behavior were investigated:

- **Condition 1:** the *uni-modal (speech-only)* condition; only verbal instructions, no gesture or gaze behaviors
- **Condition 2:** the *multimodal (speech-gesture)* condition; verbal instructions with gaze and pointing gestures

Salem et al. evaluated whether the participant's attitude towards the robot changed between the different conditions, investigating items such as perceived 'activity', 'competence', 'liveliness', 'friendliness' and 'sympathy'. All evaluated items were rated higher in the second condition, though significant differences were only measured for 'activity', 'liveliness' and 'sympathy'.

Salem's work is very similar to our own. We also investigate how non-verbal behavior influences the human-robot collaboration. But we consider an additional condition, between 1) and 2), in which verbal instructions were only supported by gaze behavior. Thereby we wanted to find out whether gaze is already sufficient to enhance the human-robot-interaction in a way the *multimodal* condition does. Furthermore we expanded the evaluation by verifying not only the participants' view of the robot but also their self-assessment in the scenario. Similar to Giuliani and Knoll we also considered objective data, such as the duration of the interaction in our evaluation.

3 Development of an Instructing Robot

Due to its role as an *instructor*, our robot NAO[1] had to offer the human participant an exact instruction as he or she is not able to guess how the jigsaw has to look at the end of the task until the last piece of the puzzle is placed. In our game puzzle pieces are colored squares with colored shapes (circle, cross etc.; also see figure 2) in their center that can be easily referred to by the robot and as well expressed by Text-To-Speech (TTS).

We implemented the abstract jigsaw on a Microsoft Surface[2] Touch-Table, from where the robot obtains the exact coordinates of the jigsaw pieces via WLAN. The robot uses this data to calculate the head orientation for the gaze direction and the arm position for pointing gestures. The robot is able to establish eye contact with the human participant by using the built-in face detection module of Aldebaran Robotics.

Each round of the puzzle is divided in a piece task and a field task. Every task uses the behavior process explained by Ishiguro [6]:

Piece Task. First the robot (R) establishes eye contact with the participant (P) (see figure 1), using only the head and no eye movements. Thereon it describes the puzzle piece the participant has to select, by a verbal utterance such as for example *"Please select the black piece with the red circle in it"*. Meanwhile it gazes and points at the mentioned piece. After that it establishes again eye contact with the user signaling that it is now the participant's turn to continue.

[1] `http://www.aldebaran-robotics.com/`
[2] `http://www.microsoft.com/surface/`

The human participant then has to react by touching shortly the searched piece with his fingers on the touch table. When a piece was touched on the Microsoft Surface, the robot changes its gaze towards the corresponding position. If it was the right piece, the robot gives the participant positive, verbal feedback while performing a small confirming head nod (= backchannel signal as in [1]). Otherwise if it was the wrong piece, it shakes its head to indicate that the participant chose the wrong piece. In this case the robot repeats the whole piece task thus offering the participant another chance to succeed the task.

Field Task. The Field Task is analog to the Piece Task as can be seen in figure 1. Except that, in this case, the robot points to the position of the puzzle field, where the participant has to drag and drop the previously selected piece. The verbal instructions describe the target position in reference to a nearby piece that is already on the field, such as "Now place this piece left to the black piece with the red circle in it".

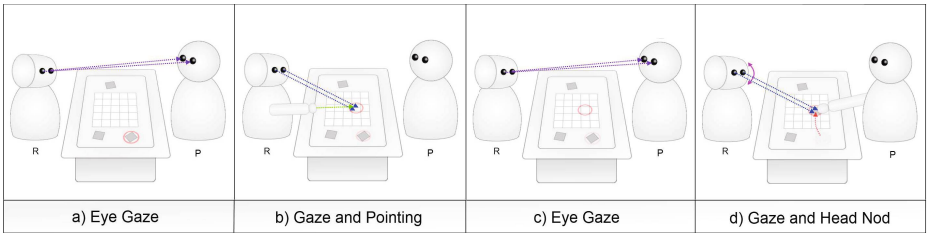


Fig. 1. Structure of a Field Task using speech, gaze and pointing gestures; Robot (R) and Participant (P)

4 Experimental Design

To compare the effects of the different levels of modality we created three scenarios with which the participants were confronted:

4.1 Scenarios

1. **Verbal utterances only:** Robot only gives verbal instructions without eye contact, gaze or pointing gestures
2. **Verbal utterances with gaze:** Robot gives verbal instructions with eye contact; uses gaze demonstrating the human participant which puzzle piece has to be placed at which position
3. **Verbal utterances with gaze and pointing gestures:** Robot gives verbal instructions with eye contact; uses gaze and pointing gestures demonstrating the human participant which jigsaw piece has to be placed at which position

4.2 Conditions

We also varied the difficulty of the puzzle game in each scenario. The varying conditions should allow further insight into the effects of the used modalities, regarding the objective performance and subjective experience of the participant. While the participants might perform comparably good under simple conditions in all scenarios, the experience of the interaction might differ significantly. And to which extent will this change when the task gets more challenging?

So we prepared two different conditions, a *simple* puzzle game and a *complex* one. The initial positioning of the puzzle pieces at the beginning of the different games was for all participants the same.

1. Simple: The simple puzzle game contains 10+1 puzzle pieces. The first puzzle piece already lies at its right position in the puzzle field enabling an easier description of the remaining pieces' positions. The remaining ten pieces are all needed to solve the puzzle game. The puzzle pieces can by unambiguously identified by their color and the shapes in their center.

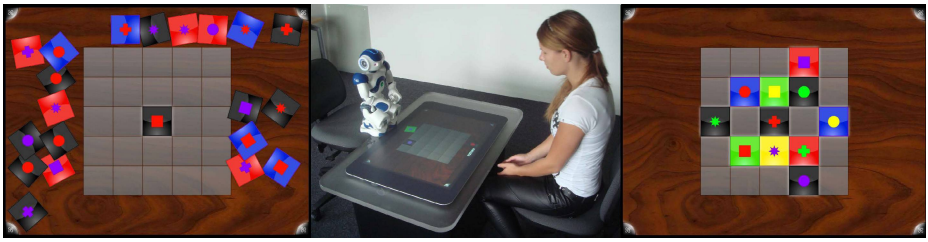


Fig. 2. Left: initial positioning of the *complex* puzzle game; **Middle:** experimental environment; **Right:** completed *simple* puzzle game

2. Complex: The complex puzzle game contains 20+1 puzzle pieces. But only ten of the remaining pieces are actually needed to solve the puzzle. The superfluous pieces are installed to complicate the detection of the relevant pieces. Furthermore there now are identical pieces, that may appear several times. Although those pieces are identical with regard to their appearance, they are handled by the robot as different pieces.

In this condition the instructions of the robot can be ambiguous, especially in the verbal scenario, as the robot describes pieces only by there colors and shapes, which are not unique anymore. Furthermore in the second condition, the gaze might not be accurate enough to distinguish identical pieces that are lying close to each other.

4.3 Participants

In the study, a total of 60 participants (9 female, 51 male) participated in the experiment, ranging in age from 19 to 56 years ($M = 26.05$, $SD = 6.24$).

All participants were recruited at Augsburg University, whereby the majority of them were students (56 of 60), mostly in Computer Science (44 of 60). Participants were randomly assigned to the different experimental scenarios.

4.4 Procedure

Participant and robot sat directly opposite each other at the Microsoft Surface touch table (see Fig. 2). First the participants got a brief introduction about the evaluation procedure. They were told they had to follow the robot's instructions to solve the abstract puzzle game and would not be able to succeed the puzzle on their own. To become familiar with the task setting and the used technologies they initially went through a small tutorial game. After the tutorial they started with the *simple* puzzle game. When they were done they were asked to fill in a questionnaire. After that they had to solve the *complex* puzzle game and fill in again the questionnaire. If an instruction was not understood, the participant had the option to touch the head of the robot to make it repeat the last instruction.

Though all participants played the puzzle game in both conditions, they were assigned to only one of the three scenarios, explained in section 4.1.

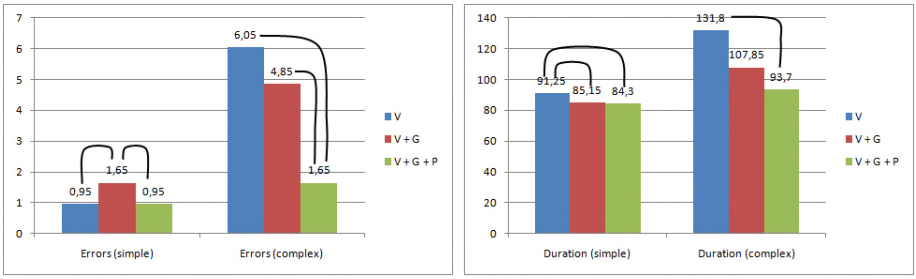
4.5 Evaluation

During the interaction the Microsoft Surface logged the errors made by the participant and how long it took to solve the puzzle, in each condition. Selecting a wrong piece, placing a piece at a wrong position and asking the robot to repeat the last instruction were counted as errors. Additional to this objective data we asked the participants whether the robot was experienced as 'attentive', 'active', 'friendly', 'lively', 'sympathetic', 'competent' and 'communicative', similar to the evaluation of Salem et al. [7]. Furthermore we wanted to know: "My experiences with Nori were better than I had expected" and "Consequences of my actions were clearly recognizable". Regarding the self-assessment of the participant we investigated seven items, including "I always immediately understood what to do next", "I think the puzzle was demanding", "I felt competent enough to comply the required tasks" and "I was completely focused on the robot's instructions".

All these items had to be rated on a five-point Likert scale with endpoints 1 = *very appropriate* and 5 = *not appropriate*. A Mixed ANOVA, a mixture of between-group and repeated-measures design, was conducted with two Within-Subject Factors *Simple* and *Complex*, summarized *Difficulty*, and one Between-Subject Factor *Scenario* (*Verbal*, *Gaze* or *Pointing*). Figures 3 and 4 illustrate the difference in the ratings. The black lines in these graphs mark the scenarios between which significant differences ($p < .05$) were revealed.

4.6 Objective Results

Errors: Figure 3(a) shows that in the simple condition only very few mistakes were made even if the participants had to rely only on the verbal instruction of



(a) Mean number of errors (b) Mean duration (in seconds)

Fig. 3. Objective data for the simple and complex puzzle games; V = Verbal, G = Gaze, P = Pointing gestures

the robot. This changes significantly in the complex condition. Our test revealed that the scenario had an significant effect on the rate of errors ($F(2, 57) = 13.68$, $p < .001$).

Planned *Helmert Contrast* illustrated that having no gestures significantly increased the number of errors ($p < .005$) and that the use of gaze and pointing gestures in comparison with applying only gaze significantly reduced the error rate ($p < .001$). Nevertheless scenarios *Verbal (V)* and *Verbal with Gaze (V+G)* did not significantly differ in the complex condition ($p > .05$) verified by the *Post Hoc Test* of *Games-Howell*.

Duration: We only consider the pure interaction time of the participant in seconds, without the time needed by the robot for its instructions, as our duration. In contrast to the average error rate a positive trend for the duration is visible in both conditions (see figure 3(b)).

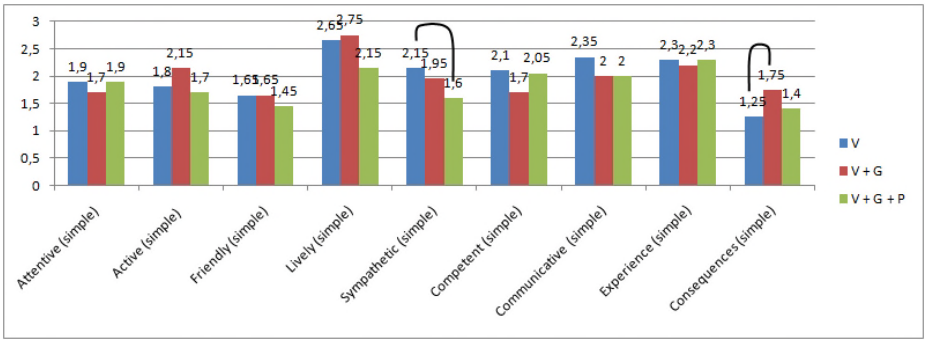
Planned *Helmert Contrast* illustrated that having no gestures significantly increased the duration of a game ($p < .005$) but only the combination of gaze and pointing gestures had a significant effect.

4.7 Subjective Results

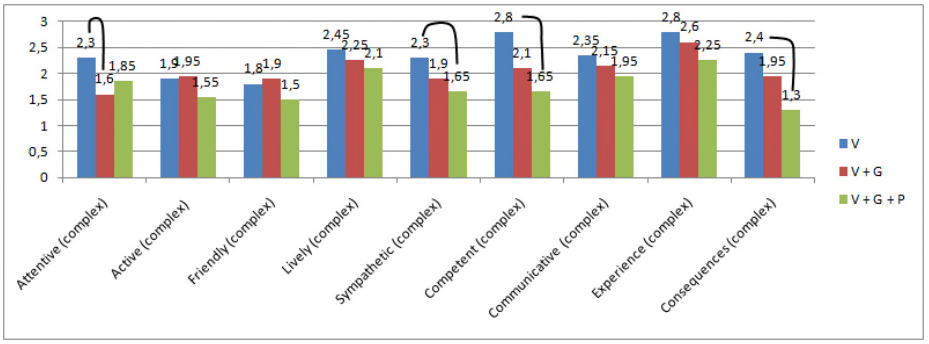
Figure 4 illustrates the mean results of our questionnaire concerning participants' perception of the robot during the interaction. For the items 'Active', 'Friendly', 'Lively' and the expected experience with the robot no significant effects could be found. On the other side significant effects were revealed for the following items:

Sympathetic: *Test of Between-Subjects Effects* yielded a significant difference between the scenarios ($F(2, 57) = 3.23$, $p < .05$). In detail, *Post-Hoc Test* of *Games-Howell* revealed that our robot using gestures is better evaluated by participants than without gestures and only speech ($p < .05$), but there is no significant difference between gaze and gaze combined with pointing gestures ($p < .38$ and $p > .05$).

Competent: There were no significant differences in the ratings for the simple condition. But a One-Way ANOVA revealed that in the complex condition the



(a) Assessment of the robot in the simple condition; lower values are better



(b) Assessment of the robot in the complex condition; lower values are better

Fig. 4. Subjective results for the simple and complex condition of the puzzle game; V = Verbal, G = Gaze, P = Pointing gestures

robot was perceived significantly more competent if it used gaze and pointing gestures ($p < .005$), than a robot that used only speech or speech with gaze.

Attentive: One-Way ANOVA revealed that in the complex condition the robot was assessed as significantly more attentive ($p < .05$) when it also used gaze in contrast to the verbal scenario. Unfortunately this effect is not significant any more when pointing gesture are added.

"Consequences of My Actions were Clearly Recognizable": *Post-Hoc Test* of *Games-Howell* verified that a robot using gaze and gestures significantly effects the participants' assessment of whether the consequences of their actions were recognizable ($p < .05$).

Regarding the self-assessment of the participants the most interesting result was that they were significantly less **focused** on the robot's instructions in the *V+G* scenario than in the *V* and in the *V+G+P* scenario both in the simple and in the complex condition (all ps $< .05$). There is no significant difference between the *V* and *V+G+P* scenario. For the item **"I always immediately understood what to do next"** there could only be found a significant improvement

($p < .005$) for the *V+G+P* scenario in the complex condition, compared to the *V* and *V+G* scenario. The mixed ANOVA also confirmed that the game was more **demanding** in the complex condition than in the simple condition (as intended), proved by the significant difference of *Difficulty* in the *Test of Within-Subjects Effects* ($F(1, 57) = 67.37$, $p < .001$). However, there was no significant difference between the scenarios, as we had actually expected.

5 Discussion and Conclusion

The main goal of this study was to evaluate how the objective performance and subjective experience of an human-robot interaction is affected by an increasing level of modality. In contrast to former studies we did not only compare the uni-modal case with the multimodal case, but we included an intermediate step and changed the difficulty of the interaction.

Regarding the objective performance our tests revealed that under simple conditions only the average duration of the interaction profits from additional modalities, compared to just verbal instructions. However the results also tell us that gaze alone is in this case not enough to make a significant difference. Quite contrary to the complex condition where a robot using speech with gaze behavior alone already achieved significantly shorter times for solving the puzzle. Adding pointing gestures improved the duration even more, but not to a significant extent. But considering the total number of errors it is definitely best to combine speech with gaze and pointing gesture, than just rely on additional gaze behavior. The participants made significantly less errors in the complex condition when they were guided by the robot with the highest level of modality. Gaze alone wasn't enough to improve the situation significantly.

This is also supported by our subjective data. In the study of Salem et al. [7] the robot in the multimodal condition (with gaze and pointing gestures) was perceived as more *active, lively* and *sympathetic*. While our robot was also more *sympathetic* to the participants when it used its full range of modalities, we could not achieve this for *active* and *lively*. For feedback our robot nodded and shook its head and sometimes moved its arms even in the *Verbal* scenario. The movements just didn't contribute to the grounding process and it didn't try do establish eye contact. This was different to the setting of Salem et al., where the robot was stiff in the uni-modal condition. The rather positive and balanced ratings in our scenarios suggest, that multimodal feedback is already enough to make a robot appear *active* and *lively*.

In the complex condition our robot was perceived as significantly more *competent* in the *V+G+P* scenario. Still the difficulty did not significantly affect the ratings, so the significant difference in the complex condition results from the poor ratings of the other scenarios. This supports that it is very important that the robot uses its full potential of modality, otherwise the perceived *competence* will drastically decline when the tasks get tougher.

Unexpected were the results that the participants were significantly less focused on the robot's instructions in the *V+G* scenario, compared to the other

scenarios. Many participants of the $V+G$ scenario stated they didn't immediately realize that the robot was moving its head and trying to establish eye contact. When the robot used pointing gestures they were aware that the robot was moving and presumably payed more attention, but still not more than when the robot was just giving verbal instructions. These observations emphasize that the nature of a task affects the effectiveness of modalities and has to be considered when designing human-robot interactions. Considering that in the simple condition more errors were made when the robot used gaze without pointing gestures, it seems that the human interaction partner might even be negatively affected (perhaps by distraction or confusion) if pointing gestures are missing.

In summary it can be said that positive trends in favor for the use of speech in combination with gaze and pointing gestures are visible and can be statistically supported for the objective performance as well as for the subjective experience of the participants. Often adding gaze behavior alone doesn't improve the interaction enough. It might even result in unexpected negative effects.

So far the participants had the role of subordinates, that never had the chance to refuse to follow an instruction. Future work will allow participants to contribute to the interaction also in different ways, to allow further insight how the use of signals, not limited to gaze and pointing gestures, affects the performance and experience of collaborative human-robot interactions.

Acknowledgments. This research was funded by the EU project Tardis (FP7-ICT-2011-7, grant agreement no. 288578).

References

1. Breazeal, C., Kidd, C.D., Thomaz, A.L., Hoffman, G., Berlin, M.: Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In: Proc. IROS (2005)
2. Burghart, C., Gaertner, C., Woern, H.: Cooperative solving of a children's jigsaw puzzle between human and robot: First results. In: Proc. AAAI (2006)
3. Clark, H.H., Brennan, S.E.: Grounding in communication. In: Perspectives on Socially Shared Cognition, pp. 127–149 (1991)
4. Faber, F., Bennewitz, M., Eppner, C., Görög, A., Gonsior, C., Joho, D., Schreiber, M., Behnke, S.: The humanoid museum tour guide robotinho. In: Proc. RO-MAN (2009)
5. Giuliani, M., Knoll, A.: Evaluating Supportive and Instructive Robot Roles in Human-Robot Interaction. In: Mutlu, B., Bartneck, C., Ham, J., Evers, V., Kanda, T. (eds.) ICSR 2011. LNCS, vol. 7072, pp. 193–203. Springer, Heidelberg (2011)
6. Ishiguro, H., Ono, T., Imai, M., Maeda, T., Nakatsu, R., Kanda, T.: Robovie: An interactive humanoid robot. Industrial Robot: An International Journal, 498–503 (2001)
7. Salem, M., Rohlfing, K., Kopp, S., Joublin, F.: A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot-interaction. In: Proc. RO-MAN (2011)
8. Sugiyama, O., Kanda, T., Imai, M., Ishiguro, H., Hagita, N.: Natural deictic communication with humanoid robots. In: Proc. IROS (2007)