The Role of the Timing between Multimodal Robot Behaviors for Joint Action

Kerstin Fischer, Lars Christian Jensen, Stefan-Daniel Suvei & Leon Bodenhagen
University of Southern Denmark

Introduction & Previous Work

That timing in HRI is important is clear ever since Suchman (1987) has demonstrated the crucial role of timing for all interactions with technology. She showed that if system response is delayed, users, for whom the system behavior is not transparent, consider the lack of timely response as failure and initiate a new action, to which the system may respond with the previously initiated behavior, or it may abort the current action and start the next behavior, or the whole process may result in error altogether. Thus, users expect a timely response to their actions as a precondition for joint action.

However, besides concerning timeliness in response to human action, timing also plays a role in the coordination of the robot's own actions. This is particularly evident in embodied conversational agents; here, much previous work concerns the synchronization of speech, gaze and gesture (e.g. Skantze et al. 2014; Mehlmann et al. 2014); these studies show that multimodal integration contributes considerably to the agents' perceived 'naturalness' and 'liveliness'. However, models of multimodal processing have not been extended to the integration of speech, navigation and gesture/manipulation, i.e. actions that play a crucial role in human-robot joint action. In social robotics, much work concerns the timing of the robot's behavior with respect to the human's behavior (e.g. concerning gaze, cf. Mutlu et al. 2012, Fischer et al. 2013), yet the synchronization of robot behaviors such as movement of the body, speech and arm movement, for instance, has rarely been addressed. So what the timing should be between, say, movement, speech and gesture in order to allow for smooth joint action is sill open.

In interactions between humans, Clark & Krych (2004) have investigated how individual actions, such as holding or placing on object, function as communicative acts. While they don't focus on timing, they show that speech and object placement are generally very well coordinated in order to allow the partner to infer the other's intentions and to predict the next move (cf. also Clark 2002). Thus the appropriate timing of multimodal action leads to legibility of this behavior and thus contributes to predictability of the actor. To investigate the role of timing of multimodal robot behavior, we carried out an experiment in which the robot either employed its multimodal behaviors sequentially or synchronized and analyzed the effects of the timing on joint action.

Method

To determine the effects of multimodal actions compared to sequential actions we elicited 36 interactions between naïve users and a large service robot.

The Robot

The robot used for the experiments was a Care-O-bot 3, a so-called 'welfare robot', developed at Fraunhofer IPA (Graf et al. 2009). The robot is approximately four feet and seven inches tall, moves on four wheels and is equipped with a robotic manipulator (arm) with seven degrees-of-freedom. The manipulator is equipped with a three-fingered dexterous hand with tactile sensors as a well as a tray that also works as a touch screen. For use in the current experiment we attached a plastic cylinder to the arm. When activated a small LED emanated from inside the cylinder. This small "device" worked ostensibly as a blood pressure measuring device.

Experimental Conditions

The study uses a between-subject design with two experimental conditions. In one condition, the robot says: "Please put your finger into the sensor so that I can measure your blood pressure", where after the robot drives up to the participant and extends its arm. These actions are performed sequentially, so that when one action ends, the next one commences. In the second condition these three modalities (speech, movement and gesture) are performed simultaneously, and are thus more fluid.

Participants

Participants were recruited from the University of Southern Denmark. We recruited primarily students and staff from the university, but also people from the general public during a public event. Mean age of participants is 30.7 with a standard deviation of 9.9. 24 of the participants are men, while the remaining 12 are women. Participants were paid with chocolate as compensation for their time and participation.

Procedure

Participants met with the experimenter outside the lab, where they signed a consent form and had their picture taken and then shown into the lab where they met the Care-O-bot. They were then led into a room to fill out a survey. Subsequently they were shown around our department and introduced to our two Keepons so that about 30 minutes passed before they were taken back to the Care-O-bot where the reported experiment took place. After the experiment they filled out a questionnaire and were offered chocolate as a compensation for their time.

Analysis

The focus of the current investigation is on participants' behavioral responses to the robot's actions and is analyzed by using ethnomethodological conversation analysis of video recordings. The analysis is supplemented with a quantitative analysis, which was done by counting the visible signs of confusion and insecurity shown by the participants, in particular: whenever people stepped back, looked searchingly at the robot, hesitated or when they asked what to do.

Results

The qualitative analysis reveals that the timing of individual actions plays a crucial role in how people understand how they should interact with the robot. The lack of synchronization of robot speech, movement and action in the first condition lead to confusion and insecurity on the part of the users. This is demonstrated in the following excerpt. The robot approaches the person and requests that she puts her finger into a sensor that is not yet visible. However, she reacts by pointing her index finger into the robot's "eyes" shortly before the robot finishes its utterance. As the robot then pulls out its arm, to which the "sensor" is attached, she realizes her mistake and now correctly puts her finger into the tube.

At the time of the verbal instruction the robot's arm movement is not completed, yet the participant responds immediately to the verbal instruction, looking for what is available to her at that moment. Her first attempt to comply with the instruction thus fails, which leads to a repair sequence and another attempt: once the robot's arm movement is complete, she identifies the right sensor while still being uncertain.

Excerpt 1

Robot says:
"please put your finger into the sensor"

Robot extends arm with sensor

here?
like that?

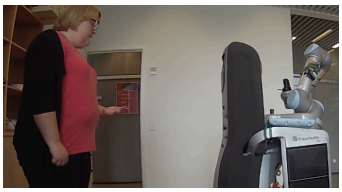
no?
it's not the right one

here?

A similar reaction can be observed in excerpt 2. Here, the participant first displays puzzlement over what to do. Similar to the participant in excerpt 1, her first attempt to comply fails and leads to a repair sequence. She then orients towards the arm as it reaches its set position and successfully complies with the robot's request.

Excerpt 2

Robot says:
"please put your finger into the sensor"



The participant takes out her hand and looks at it

Robot extends arm with sensor

She then points her finger directly at the robot

Ahh! (participant redirects her finger at the arm being extended)

Excerpts 1 and 2 demonstrate how participants make false predictions when the robot's actions are not synchronized. Similarly, in excerpt 3, the participant displays readiness to act as soon as he hears the word "sensor". He does this by moving his hands up and gazes to both sides of the robot. Meanwhile, the robot finishes its utterance and then sets the arm in motion. However, the participant experiences a significant pause (3.5 seconds) from when he displays an understanding of the request and until he shows that he knows how to comply with it.

Excerpt 3

Robot says:
"please put your finger into the sensor"

Robot extends arm with sensor

The participant looks searchingly to both sides of the robot

Ahh! (participant smiles while looking at arm being extended)

In contrast, when the robot uses all of its modalities simultaneously, participants are much better at predicting the joint action, which is shown in the following excerpt:

Excerpt 4

The robot starts to drive, says *"please put your finger into the sensor",* and extends its arm with the sensor

The participant holds up his finger as soon the robot has made its verbal request

The participant puts his finger into the sensor the moment the robot is close enough

Here, the robot makes the verbal request, starts to drive and extends its arm all at the same time, so that all three modalities have been executed once it reaches the participant. However, the participant realizes what he is supposed to do already halfway through the robot's approach. Thus, the multimodal robot behavior is

more legible. Interestingly, in excerpt 4, the participant exaggerates his actions for the robot to take his response into account.

Results from the qualitative are supported by the quantitative analysis. The analysis shows that people interacting with the robot using inappropriate timing show significantly more signs of confusion and insecurity ($p<.05$, Fisher's Exact Test).

Discussion & Conclusion

The results show that the timing of robot multimodal actions play a crucial role even for a situation that requires as limited joint action as the one under consideration. In particular, we found that

- people assume that they should be able to carry out an instruction in the moment the instruction is uttered; thus speech needs to be carefully coordinated with the moment all preconditions for the human partner to carry out the required action are fulfilled;
- people process the robot's behavior incrementally and on the basis of partial information and start predicting the robot's actions on the basis of what is available at each given moment. This can lead to inappropriate proactive behavior if the robot's individual actions are not sufficiently legible.

This has consequences for the legibility of robot action since users don't wait for the whole action before making their predictions and acting proactively. The design of legible robot behavior thus needs to take the timing between actions and processing in time into account if human-robot joint action is supposed to be successful.

Acknowledgements

We wish to thank Sibel S. Isikli-Kristiansen and Kristina Holz for their help during the experiments. Furthermore, we gratefully acknowledge the support of the *SPIR* project *patient@home*.

References

Clark, H.H. (2002): Speeking in Time. *Speech Communication* 36: 5-13.

Clark, H. H. & M. A. Krych (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language* 50 (1), 62–81.

Fischer, K., Lohan, K., Nehaniv, C. & Lehmann, H. (2013). Effects of Different Types of Robot Feedback. *International Conference on Social Robotics'13,* Bristol, UK.

Graf, B. (2009): An Adaptive Guidance System for Robotic Walking Aids. *Journal of Computing and Information Technology* - CIT 17, 1: 109-120.

Mehlmann, G., Häring, M., Janowski, K., Baur, T., Gebhard, P. & André, E. 2014. Exploring a Model of Gaze for Grounding in Multimodal HRI. *ICMI 2014*: 247-254.

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., & Ishiguro, H. (2012). Conversational Gaze Mechanisms for Humanlike Robots. *ACM Transactions on Interactive Intelligent Systems* 1, 2, art. 12.

Skantze, G, Hjalmarsson & Oertel, C. 2014. Turn-taking, feedback and joint attention in situated human-robot interaction. *Speech Communication* 65: 50-66.

Suchman, L. (1987): *Plans and Situated Actions*. Cambridge: Cambridge University Press.