How Should a Robot Interrupt a Conversation Between Multiple Humans

Oskar Palinko(✉), Kohei Ogawa, Yuichiro Yoshikawa, and Hiroshi Ishiguro

Osaka University, Osaka 560-8531, Japan
palinko@irl.sys.es.osaka-u.ac.jp

Abstract. This paper addresses the question of how and when a robot should interrupt a meeting-style conversation between humans. First, we observed one-to-one human-human conversations. We then employed raters to estimate how easy it was to interrupt each participant in the video. At the same time, we gathered behavioral information about the collocutors (presence of speech, head pose and gaze direction). After establishing that the raters' ratings were similar, we trained a neural network with the behavioral data as input and the interruptibility measure as output of the system. Once we validated the similarity between the output of our estimator and the actual interruptiblitiy ratings, we proceeded to implement this system on our desktop social robot, CommU. We then used CommU in a human-robot interaction environment, to investigate how the robot should barge-in into a conversation between multiple humans. We compared different approaches to interruption and found that users liked the interruptibility estimation system better than a baseline system which doesn't pay attention to the state of the speakers. They also preferred the robot to give advance non-verbal notifications of its intention to speak.

Keywords: Human robot interaction · Conversational turn-taking
Social robotics

1 Introduction

Humans are very able communicators. We can partake in conversations while effortlessly managing many variables: roles, pauses, interruptions, etc. Having conversations with multiple people is a very essential human ability. For this reason, we think it would be very important for a robot to be able to participate in such interactions too, see Fig. 1(a). This involvement should be very efficient and courteous. To achieve this, the robots should be able to estimate what is a good time to barge-in when it has something to say. Not only the timing is important but also how this interruption would occur. Addressing these issues could lead to a wider acceptance of robots in conversational environments. Neglecting these issues could lead to robots being ignored or even excluded from human conversations. By studying this topic, we could not only design better robots but also advance the understanding of human behavior.

S. S. Ge et al. (Eds.): ICSR 2018, LNAI 11357, pp. 149–159, 2018.
https://doi.org/10.1007/978-3-030-05204-1_15

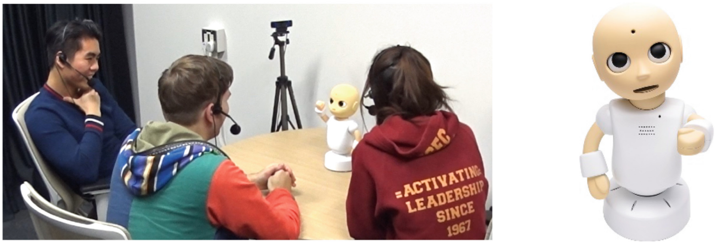

Fig. 1. (a) Multi-party conversation with a robot. (b) CommU, desktop robot.

2 Background

Turn-taking is an essential approach to understanding many human activities, including multi-party conversations. Sacks et al. defined a widely used model for conversational turn-taking [1]. Gaze also plays a critical role in selecting the next speaker in such a scenario [2]. Very often the next in turn corresponds to whomever the previous speaker was looking at last.

According to [3, 4] a period of silence in conversational turn-taking signals the relinquishment of the turn by the previous speaker. This means that anyone who wishes to barge-in is welcome to take over the conversation.

Spoken dialog systems are very good examples of turn-taking [5]. Barge-ins are common events in such systems [6]. This is when a user interrupts an automatic voice system to reduce the waiting time. In our research the barger-in would not be the person, but rather the robot. None the less, there are lessons to be learned from the human experience of barging-in.

Typically, human-robot conversation studies involve one robot and one human. For example, Snider et al. explored what influences engagement in human-robot conversations [7]. They found that gestures are important for creating a more engaging environment. Other specific HRI research setups might involve two or more robots. Hayashi et al. looked at how two robots talking to each other might influence a human listener. They found that robots can be effective passive social agents [8]. Iio et al. investigated how multiple robots can improve conversation initiation with a human [9]. They found that attracting people's attention works better with two robots than with one. On the other hand, not many studies looked at a robot interacting with multiple humans in conversations. Mutlu et al. researched how eye gaze can influence the interaction between a robot and two people [10]. Even though the conversation was kept almost the same, gaze could effectively influence people's attitude towards the robot. Matsusaka et al. designed a robot system which could interrupt a conversation between two people when it had corrective information to provide [11, 12]. It used multi-modal sensing (voice, face, speech, etc.) to determine when to barge-in. The authors did not validate or test their system in an experimental study in these papers.

Interruptions are essential parts of human-computer interaction. Coordinating these is a non-trivial task, which warrants in-depth study [13]. Interruptions arriving at wrong times can lead to decrease in human performance [14]. Therefore, it would be important to know the level of interruptibility of people in conversations and other tasks. Stern et al. define interruptibility as "the current state of a user regarding her receptiveness to receive messages" [15]. In other words, high levels of this measure indicate a person who is very able to receive interruptions, while low levels mean that the person is very involved with another task and should not be bothered. This measure is also very important in human-robot interaction [16–19] for robots to understand when they should talk to humans. Banerjee and Chernova designed a discreet value estimator of interruptibility for a mobile robot [20]. It was able to tell how interruptible people were in its environment. They did not focus on conversations between people and robots. Their work included validation of the estimator but not an experimental study of the designed system.

3 Approach – Interruptibility Estimator

Instead of coming up with our own rules on when interruptions should occur we decided to take the learning approach: learn from humans when it is good to interject a conversation. In this sense we adopted the concept of interruptibility: the higher the value, the easier it is to interrupt the person. We decided to use a regression-type artificial neural network (ANN) for learning this measure, a multilayer perceptron. The inputs to the network would be signals which can be observed by our senses during a conversation, e.g. speech, head pose and eye gaze, while the output would be the level or interruptibility. By speech, we mean the presence or absence of a speech signal (on/off) which could be determined using a microphone with a given threshold. Of course, there is more information embedded in speech itself in terms of linguistic analysis which could help to determine how interruptible someone is, but here we chose to focus on simple signals, as presence of speech. The reason for this simplified approach is that today's state-of-the-art speech recognition systems include a considerable amount of delay, which could undermine the ability of a robot to interrupt a conversation quickly. Also, only speech recognition would not be enough. Rules should be devised based on language models which would allow a conversation to be interrupted. This would add even more complexity and delay.

The output signal of the ANN was the level of interruptibility. It is a subjective measure, thus we collected ground truth data from several raters and checked for their agreement, similarly as in [20]. We decided to use one hidden layer of 15 neurons with hyperbolic tangent activation function so that the system could account for nonlinearities which might occur.

4 Learning and Validation Experiment

We asked 6 pairs of students from our lab to have one-to-one conversations about topics of common interest. They were seated in a quiet room facing each other. Separate videos of each of the subjects were recorded with sound coming from directional headset microphones. Each interaction lasted for about 2 min and 15 s. Other than talking, we asked the participants to spend around 15 s in silence while writing something down on a piece of paper in front of them. We also asked them to look up something on their smartphones for another 15 s. These additional tasks were added to create a varied input signal for the ANN. Once the videos were recorded the experimenter annotated the conversations to determine when there was speech. We also ran a face pose [21] and eye gaze [22] detection algorithm and recorded the data. These three groups of signals (audio, head pose, gaze) were then used as the input for our estimation system. As we used a supervised learning approach, we needed to obtain the output signal ground truth. For this, we divided the video recordings of the conversations into one second long clips. The step was 0.5 s, so two subsequent clips had 50% overlap. The first author and two hired raters watched the clips of all subjects, giving an interruptibility rating at every 0.5 s. They were asked to rate the interruptibility at the end point of each clip. The ratings were integer numbers from 1 to 7. The value ‘1’ meant: Only interrupt with an emergency message like ‘the house is on fire’. The value ‘7’ meant: easily interruptible with any message. All the other values were evenly spaced between the two extremes. The two raters’ ratings were used to check for inter-rater reliability of the first author’s values. Once the ratings were obtained we calculated Cronbach’s alpha, which resulted in $\alpha = 0.826$. This means good inter-rater reliability, so we continued with using the experimenter’s interruptibility ratings for training the ANN, similarly as [20].

The input signal of the ANN consisted of audio presence and its past 9 values, head pose (roll, pitch, yaw angles) and gaze (roll, pitch, yaw angles). The output layer used linear activation. We tried including past values of head pose and gaze too, but they did not make a difference in results.

Once the estimator was trained, we proceeded to evaluate the quality of estimation. To do so we used the leave-one-out approach: we trained the system with the data of 11 subjects and then tested on the 12th person. We repeated this 12 times once to test on each subject’s data. We recorded the output of the estimator in each case. Then we compared the similarity of the estimator’s output to the original interruptibility rating of the experimenter, again using Cronbach’s alpha. We compared using only audio data as input to using additional signals as head pose and eye gaze in addition to the audio data. The results can be found in Table 1. For both situations Cronbach’s alpha indicated a good agreement between the original and estimated values. It should be noted that adding head pose and eye gaze improved the output of the estimator, proving that these additional signals can be used for enhancing interruptibility estimation.

Table 1. Validation results between ground-truth and ANN output.

	Cronbach’s alpha
Audio only	0.804
Audio, head pose, gaze	0.815

5 Exploratory Human-Robot Interaction Study

Once we successfully verified our interruptibility estimator, we proceeded to test it in an actual human-robot interaction study. The purpose of the study was to explore in what exact ways a robot could interrupt a conversation between humans with minimal negative effects.

5.1 CommU, the Desktop Social Robot

The robot in question was CommU, a 30 cm tall desktop humanoid robot, see Fig. 1 (b). It is a commercially available robot with a fixed platform and quite sophisticated movements (for its size) in its torso, arms, head, eyes, mouth and eyelids. Its strong points are its oversized eyes which are capable of both gradual and fast movements, thus emulating smooth pursuit and saccades of the human eye. It has a built-in speaker and its mouth emulates human mouth movements while it speaks using a text-to-speech engine.

5.2 Experimental Setup

Our experimental setup included a round table (d = 105 cm) with CommU on top of it, two webcams, three chairs and three subjects with headset microphones. Figure 2(a) shows the layout.

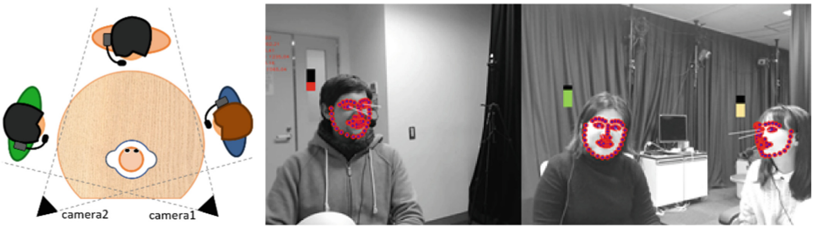

Fig. 2. (a) Experimental setup. (b) Face, gaze and interruptibility tracking.

The cameras were set up to track the faces of the three participants. We used two Logitech C920 webcams to cover a 180-degree field around the robot. We used the CLM algorithm to track faces [21] and the approach in [22] to track gaze direction, see Fig. 2(b). This information was directly fed into the interruptibility estimator algorithm. The audio signal of each participant was captured using a directional headset microphone. After thresholding, the microphone signal was also connected to the input of the estimator system. The robot was shifting its gaze from face to face in a random order, but when somebody started speaking, the robot looked at the speaker. If more than one person was speaking, the gaze was shifted towards the loudest one.

As there were three subjects at a time we created one estimator for each of them. The condition was to have the interruptibility level for all three persons above the middle value of 4, for the robot to start speaking or raise its hand.

5.3 Experimental Conditions

The only independent variable of the experiment was the strategy used by the robot to interact with people. The following were the three levels of the variable:

(1) *direct speech* – the robot starts speaking immediately after receiving a command
(2) *wait to speak* – the robot passively waits until the interruptibility levels for all participants rise above the middle value of 4 then speaks,
(3) *hand up* – the robot waits for interruptibility to rise as in the previous condition but does not do so passively. Rather, it raises its hand right away after receiving the command to signal that it has something to say, see Fig. 1(b). After this, it waits for the levels to rise. If that doesn't happen, after 5 s it says, "Excuse me" and waits again. If the levels do not become satisfactory again, after 10 s it says, "Can I say something?" After this, it continues waiting for interruptibility to reach appropriate values, passively. These three conditions were selected to test a wide range of robot behaviors.

5.4 Experimental Procedure

We recruited 18 participants through university advertising (11 male and 7 female). Their average age was 24.3 years. As we needed 3 people for the conversation, we invited them three by three. Most of the subjects within a group were familiar with each other. All groups were exposed to all three conditions. The order of the 3 conditions was counter-balanced between groups. Each of them received a participation reward equivalent of about 18USD. All participants were naïve towards the purpose of the experiment and have not interacted with CommU before.

After the subjects arrived they were given a description of the experiment and asked to sign consent forms. An initial questionnaire gauged their affinity towards robots ("How open are you towards robots in general?", "How open are you towards robots participating in conversations?") and asked for their basic data (name, age, etc.) They were asked to have a natural conversation between each other and let the robot join in as much as possible. Other than this we didn't tell them anything else about what kind of behavior they could expect from the robot and how to interact with it. We ensured they stay naïve towards the goal of our study. They were assured that the robot will not make any unexpected actions and that it cannot harm them. In each experiment the subjects were asked to have a conversation about these topics: (1) food you would suggest visitors to try in Japan, (2) places to visit in Japan, (3) how to learn to speak Japanese. The topics always followed this order (food first, travel second, language third), while the interaction conditions were counter-balanced in order. Each session lasted for about 15 min with a short break after each.

After the experiment ended, participants were asked to fill out a post-experiment questionnaire, checking if their affinity for robots changed and asking them to give their opinion about four statements. The statements were: (1) the robot could participate in the conversation, (2) I was satisfied with the robot's behavior, (3) the robot was very polite in this session, (4) I would like to have a robot with this behavior in a meeting/conversation. The offered answers to these questions were similar as for interruptibility rating: evenly spaced integers from 1 to 7.

During the sessions the experimenter provided robot utterances (unbeknownst to the subjects) from a predefined set of sentences, similarly as in [10]. About half of these were non-topic-specific ("I completely agree", "I'm afraid I disagree with that.", "Interesting! Could we get anyone else's opinion?"), while the other half were topic-specific for each session (Session A: "I like sushi. How about you?", "Have you tried soba noodles?", etc. Session B: "Have you been to Hokkaido?", "I hear the ocean is really beautiful in Okinawa. What do you think?", etc. Session C: "What book would you suggest for a beginner?", "Does the university offer free Japanese classes?", etc.) The robot utterances were dispatched by the experimenter only at times when people were actively engaged in conversation. This was done because we wanted to study how conversation interruptions could be managed by the robot and how people would react to them. If we would have generated utterances during times without conversations, those would be wasted, because they do not need any strategy to deliver: they could be said right away in all conditions because the interruptiblity would be high.

For the first two groups of subjects the *hand up* condition consisted of the robot raising its hand without saying anything and waiting for the interruptibility level to rise for all participants before uttering the sentence. But both groups of subjects failed to recognize this signal as a sign of the robot wanting to speak, thus we realized that the signal needs to be emphasized for it to be successfully interpreted. This is why we changed it for the subsequent four groups as follows: hand is raised and interruptibility rise is expected. If that doesn't happen, after 5 s the robot says, "Excuse me" and waits again. If the levels do not become satisfactory again, after another 5 s it says, "Can I say something?" After this, it continues waiting for the levels to reach appropriate values, passively. This sequence of actions was much better recognized by participants as a request to speak (100% recognition rate). In the following section we report on the results generated by the last four groups of subjects (12 people in total) who performed the *hand up* condition in the same enhanced way.

6 Experimental Results

An important measure of how efficiently the system performed is the delay between when a command was sent to the robot and the time when the robot uttered the sentence. For the *direct* condition this time was by definition equal to zero, because the robot was programmed to say the received utterance without delay. In the *wait* and *hand up* conditions CommU waited until interruptibility rose before speaking. Figure 3 (a), shows how the delay compares for the two conditions.

The difference between these two delays was found to be statistically significant using a paired t-test with $t = 3.47$ and $p < 0.05$. In the *wait* condition, the delay times were quite long because the robot did not give any signal to the humans that it wanted to speak. On the other hand, in the *hand up* condition the robot clearly signaled its intention to speak by first raising its arm and then politely uttering if needed. Clearly the *hand up* option performed much more efficiently, saving a lot of time.

In the *hand up* condition 53% of the time the robot had only to raise its hand and was allowed to talk. Another 35% of the times the robot had to add "Excuse me" and only

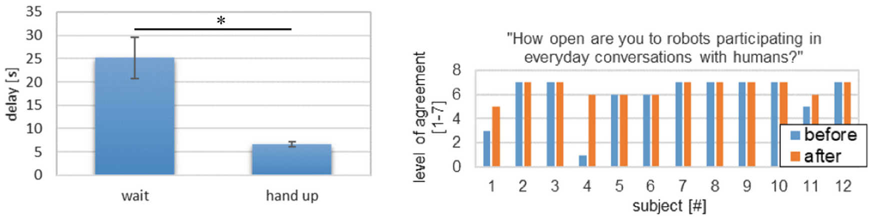

Fig. 3. (a) Time delay to utterance ± 1 std. error. (b) Before/after question.

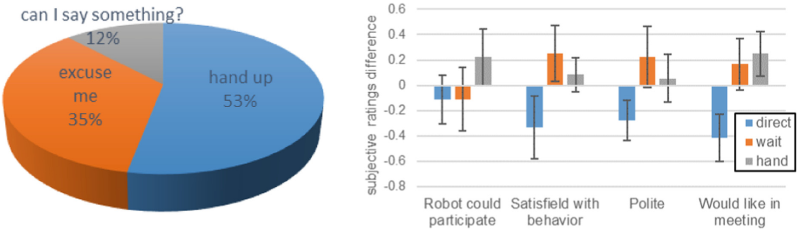

Fig. 4. (a) Robot action types. (b) Subjective opinions ± 1 std. error.

12% of the time it had to say the second sentence too “Can I say something?”, see Fig. 4 (a). This meant that people realized quite easily that the robot is asking for attention.

We gauged the participants’ opinion of robots in general and robots in conversations by asking them the same two questions before and after the experiment. The results can be seen in Fig. 3(b). It can be noticed that one quarter of the participants experienced an increase in their opinion of robots during our experiment. Three quarters maintained their opinions, while no one thought worse of robots after interacting with CommU. The majority of those who maintained their opinions already had a top mark at the beginning of the experiment, so they could not increase it at the end.

Finally, we report on the results of the experiment questionnaire. Even though none of the differences were found to be statistically significant, we still decided to include them to show the trends of opinions as this was an exploratory HRI study.

To be able to easily compare opinions, we subtracted the means of the three conditions for each subject and each question before averaging them together. This operation is warranted as we are interested in the relative differences between the three conditions and not their absolute values.

When asked the question if “The robot could participate in the conversation well”, the *hand up* condition tended to score higher compared to the other two, Fig. 4(b). When asked if they were satisfied with the robot’s behavior, participants were dismissive towards the *direct* condition and liked more the other two. In the same figure, we can notice that they rated the *wait* option as most polite, *direct* as least polite while *hand up* was in between. When asked the question “Would you like to have this system in a meeting?” subjects disliked the *direct* method while rating the *hand up* condition as best.

7 Discussion

Looking at the time delay data, Fig. 3(a), we can see that the *wait* condition caused long delays. This was because subjects were not informed about the robot's intention. There was also a lot of variation in this condition because the robot had no control over people's behavior. In the *hand up* condition, the delays were much shorter and less varied, because people received a signal that the robot wants to speak. At first, some of the subjects could have been confused what the robot wants with the raised hand (as in the first two groups which were excluded from the analysis), but if they didn't start paying attention to CommU, it would say "Excuse me" and "Can I say something?" These were very explicit requests for letting the robot take its turn in the conversation. This behavior turned out to be much more efficient in terms of delay, but it could also be interpreted as somewhat intrusive, as the robot could raise its hand at any time, with no regard to the state of the conversation. The addition of this behavior very significantly reduced the waiting time in the conversation compared to the *wait* condition.

Analyzing subjective measures, we have seen in Fig. 4(b) that people's opinion of robots either increased or stayed the same as before the experiment. Even though this measure might be biased by the participants' potential kindness towards the experiment, it is still noteworthy that nobody reported a decrease in satisfaction.

As mentioned before, subjective results were not found to be statistically significant, but we still think reporting on the trends of opinions might be beneficial, because of the exploratory nature of our experiment. As the *direct* method received the worst ratings from participants compared to the other two reactive approaches we think it justifies the need to detect and try to adapt to people's level of interruptibility.

8 Conclusion and Future Work

In this study, we set out to explore how and when robots should interrupt the conversation of multiple people in a meeting-style environment. At first, we employed a learning approach to model participants' interruptibility (the measure of how easy it is to interrupt them). As common sense would dictate, we have found that when people were silent, they were deemed to be more interruptible by raters than when they talked. But we also expanded this by signals which improved the estimation, namely head pose and eye gaze. The ANN estimator we created for this purpose was validated to be good at recognizing interruptibility. Once the modeling was done, we set out to explore how a robot equipped with this system would perform in an experimental scenario where three people would have a conversation with it. We found that the robot could barge-in into a conversation more efficiently when it first gave non-verbal signals that it wants to speak (*hand up* vs. *wait*). At the same time, the robot was deemed to be more polite and more appropriate for a conversation environment if it did give non-verbal signals as opposed to just barging-in without any notification (*hand up* vs. *direct*).

We do recognize that even though we were able to design an efficient and polite system, its performance could be improved if the robot would be equipped with a real-time speech recognition system that would allow linguistic modeling of the conversation. We are considering including this kind of improvement in future versions of the system. We also note that participants of the learning phase of this study had a direct influence on the robot's behavior. Therefore, in the future, we could influence the robot's communication style by selecting participants with desired behavior.

Acknowledgements. This research was partially supported by ERATO ISHIGURO Symbiotic Human-Robot Interaction Project and Itoki Corporation. The authors would like to thank Yutaka Nakamura for his help with learning algorithms.

References

1. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language (Baltim) **50**(4), 696 (1974)
2. Argyle, M., Cook, M.: Gaze and Mutual Gaze. Cambridge University Press, Cambridge (1976)
3. Schiffrin, D.: Discourse Markers, vol. 107. Cambridge University Press, Cambridge (1987)
4. Nagao, K., Takeuchi, A.: Social interaction: multimodal social interaction: conversation with social agents, vol. 94. In: AAAI (1994)
5. Heins, R., Franzke, M., Durian, M., Bayya, A.: Turn-taking as a design principle for barge-in in spoken language systems. Int. J. Speech Technol. **2**(2), 155–164 (1997)
6. Ström, N., Seneff, S.: Intelligent barge-in in conversational systems. In: International Conference on Spoken Language Processing, pp. 1–4 (2000)
7. Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artif. Intell. **166**(1–2), 140–164 (2005)
8. Hayashi, K., Kanda, T., Miyashita, T., Ishiguro, H., Hagita, N.: Robot Manzai - robots' conversation as a passive social medium. In: Proceedings of 2005 5th IEEE-RAS International Conference on Humanoid Robots, vol. 2005, pp. 456–462 (2005)
9. Iio, T., Yoshikawa, Y., Ishiguro, H.: Starting a conversation by multi-robot cooperative behavior. In: Kheddar, A., et al. (eds.) Social Robotics, vol. 10652. Springer, Cham (2017)
10. Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., Hagita, N.: Footing in human-robot conversations: how robots might shape participant roles using gaze cues. Hum. Factors **2**(1), 61–68 (2009)
11. Matsusaka, Y., Fujie, S., Kobayashi, T.: Modeling of conversational strategy for the robot participating in the group conversation. In: Proceedings of the 7th European Conference on Speech Communication and Technology, pp. 2173–2176 (2001)
12. Matsusaka, Y., Tojo, T., Kubota, S.: Multi-person conversation via multi-modal interface - a robot who communicate with multi-user. In: Eurospeech, pp. 1723–1726 (1999)
13. Sasse, A., Johnson, C., et al.: Coordinating the interruption of people in human-computer interaction. In: Human-computer interaction, INTERACT, vol. 99, p. 295 (1999)
14. Gillie, T., Broadbent, D.: What makes interruptions disruptive? A study of length, similarity, and complexity. Psychol. Res. **50**(4), 243–250 (1989)
15. Stern, H., Pammer, V., Lindstaedt, S.N.: A preliminary study on interruptibility detection based on location and calendar information. In: Proceedings of CoSDEO, vol. 11 (2011)
16. Mutlu, B., Forlizzi, J.: Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. In: 3rd International Conference on Human-Robot Interaction (HRI) (2008)
17. Rosenthal, S., Veloso, M.: Is someone in this office available to help me? J. Intell. Robot. Syst. **66**, 205–221 (2011)

18. Shi, C., Shiomi, M., Kanda, T., Ishiguro, H., Hagita, N.: Measuring communication participation to initiate conversation in human–robot interaction. Int. J. Soc. Robot. **7**(5), 889–910 (2015)
19. Satake, S., Kanda, T., Glas, D.F., Imai, M., Ishiguro, H., Hagita, N.: How to approach humans?: strategies for social robots to initiate interaction. J. Robot. Soc. Japan **28**(3), 109–116 (2010)
20. Banerjee, S., Chernova, S.: Temporal models for robot classification of human interruptibility. In: Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems (2017)
21. Baltrusaitis, T., Robinson, P., Morency, L.P.: 3D constrained local model for rigid and non-rigid facial tracking. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2610–2617 (2012)
22. Palinko, O., Rea, F., Sandini, G., Sciutti, A.: A robot reading human gaze: why eye tracking is better than head tracking for human-robot collaboration. In: Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)