The utility of gaze in spoken human-robot interaction

Maria Staudte
Department of Computational Linguistics
Saarland University
Saarbruecken, Germany
masta@coli.uni-saarland.de

Matthew Crocker
Department of Computational Linguistics
Saarland University
Saarbruecken, Germany
crocker@coli.uni-saarland.de

ABSTRACT

Psycholinguistic studies of situated language processing have revealed that gaze in the visual environment is tightly coupled with both spoken language comprehension and production. It has also been established that interlocutors monitor the gaze of their partners, so-called "joint attention", as a further means for facilitating mutual understanding. It is therefore plausible to hypothesise that human-robot spoken interaction would similarly benefit when the robot's language-related gaze behaviour is similar to that of people, potentially providing the user with valuable non-verbal information concerning the robot's intended meaning or the robot's successful understanding. In this paper we report preliminary findings from an eye-tracking experiment which investigated this hypothesis in the case of robot speech production. Human participants were eye-tracked while observing the robot and were instructed to determine the 'correctness' of the robot's statement about objects in view. Specifically, we examined the human behaviour in response to incongruency of the robot's gaze behaviour and/or errors in the statements' propositional truth. We found evidence for both (robot) utterance-mediated gaze in human-robot interaction (people look to the objects that the robot refers to linguistically) as well as for gaze-mediated joint attention, i.e. people look to objects that the robot looks at. Our results suggest that this kind of human-like robot-gaze is useful in spoken HRI and that humans react to robots in a manner typical of HHI.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; J.4 [**Social and Behavioral Science**]: Psychology

Keywords

gaze, joint attention, incongruency, utility

1. MOTIVATION

People have developed very subtle and complex strategies to communicate effectively, seamlessly integrating a variety of non-verbal signals during spoken language communication. Gaze as well as gestures, facial expressions and para-verbal feedback constitute some of these signals and they enrich communication in many social aspects and establish robustness. They help to convey information about attitude, emotional or belief state or simply coordinate the conversation by indicating turn-taking actions and let the partner know what the current focus of interest is. Psychological studies have revealed, for example, that gaze in the visual environment is tightly coupled with both spoken language comprehension [7, 8, 14] and production [10, 4]. It has also been established that interlocutors monitor the gaze of their partners (see e.g. [3] for a comprehensive account of joint attention). It is therefore plausible to hypothesise that human-robot spoken interaction would similarly benefit when the robot's language-related gaze behaviour is similar to that of people: not only would such behaviour imply human-like language processing, but it also provides the user with valuable non-verbal information concerning the robot's intended meaning (during robot production) or the robot's successful understanding of a user utterance (during robot speech recognition). In this paper we present work in progress and report findings from an eye-tracking experiment which investigated this hypothesis in the case of robot speech production.

Considerable work has already been done on gaze in HHI as well as robot gaze in HRI, e.g. during turn-taking [2] or with respect to information structure of the generated utterance [12]. Robot gaze generally in conversational engagement and in relation to some reference resolution has been explored by [13] among others and it could be established that the perception of robot gaze is coupled to the robot's head orientation [6]. The psychological findings from HHI, that have motivated our work, however, have not yet been applied in HRI . The role of utterance-mediated gaze in production as being tightly coupled to overall apprehension of an utterance has been established by [4], for instance. It has been shown, for example, that referential gaze is part of the planning process of an utterance and, thus, precedes the onset of the corresponding linguistic reference by approximately 800msec - 1sec. [9]. On the other hand, studies investigating gaze in comprehension, have revealed that listeners use speakers' gaze to identify a target before the linguistic point of disambiguation which clearly distinguishes utterance-mediated and gaze-mediated visual attention [5]. This study shows that gaze helps to identify possible referents of an utterance, even when the speaker's gaze was initially misleading due to the experimental setup. Subjects could establish a mapping of the speaker's gaze to their own visual scene and, thus, make use of the speaker's gaze during comprehension nevertheless. It is not clear that these insights from investigations of human cognitive behaviour can be mapped directly onto human-robot communication.

Robots differ in many ways as their physical means are distinct from ours. Robots do not possess the same amount of experience and world knowledge nor are they typically familiar with our communicative conventions. Hence, it is our general aim to investigate to what extent insights from human utterance-mediated gaze behaviour are sensibly applicable to robot gaze.

Our interest and the presented study focus on utterance-driven gaze behaviour by the robot, e.g. fixations towards an object before it is mentioned. Human gaze can then be observed in response to both the robot's speech (utterance-mediated attention) and the robot's gaze itself (joint attention). We conducted an initial experiment to show that our experimental design is generally valid and yields objective measures like decision/response times as well as the distribution of fixations to regions in the scene (which may bear evidence for other subjective/social factors). The scenario we have created is that of a robot describing a situation in blocksworld manner and simultaneously producing fixations to referenced objects. Human participants were eye-tracked while observing the robot and were instructed to determine the 'correctness' of the statement. Induced errors include incongruency of the gaze behaviour and/or errors in the statements' logical truth. These potentially reveal both the subject's attitude towards the robot as well as the utility of robot gaze in assessing validity of the robot's statements.

2. EXPERIMENT

2.1 Purpose and requirements

The presented pilot study aims to provide general empirical support for our hypothesis and method. Thus, its results provide only cues for further research trying to answer questions concerning the utility of robot gaze. If, indeed, gaze is a significant element of HRI then we can assume that *inappropriate* gaze behaviour may lead to some kind of disruption or slow-down in communication. In contrast, when the behaviour is consistent with (yet to be established) HRI conventions, we might expect interaction to be more fluent and efficient and, consequently, the acceptability and naturalness to rise. In this case, our longer term goal will be to find out what those gaze conventions are and what constitutes optimal robot gaze.

To begin investigating these issues, we require an experimental design that allows us to control the type and the occurrence of gaze and speech errors that might occur in robot speech production. Simultaneously, a method is desired that enables the experimenter to precisely observe the human subject and measure the reaction. A video-based setup fulfills these conditions by allowing the experimenter to very carefully plan and control errors and timing off-line while the subject's reaction can be recorded using an on-line eye-tracking technique. Although it might be argued that this is not real interaction, it has been shown that a video-based scenario without true interaction yields similar results to a live-scenario and can be considered to provide (almost) equally valuable insights into the subject's perception and opinion [17].

2.2 Methods

2.2.1 Participants

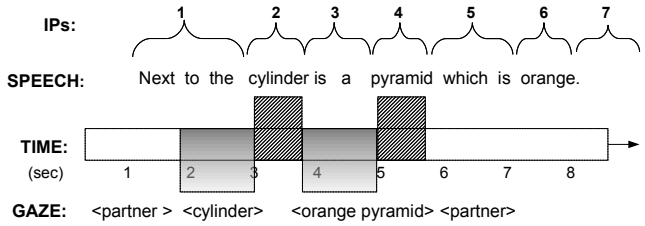
Ten students of various subjects, all enrolled at Saarland University and native speakers of German, took part in this pilot study. They had mostly no experience with robots nor with eye-tracking. They were told that the eye-tracker camera was monitoring their pupil size and, thus, the cognitive load of the task on them.

2.2.2 Material

Each video-clip showed a PeopleBot robot [1] onto which a stereo vision camera on a pan-tilt-unit was mounted, as it stood behind a table with a set of coloured objects in front of it. The objects were plain geometrical shapes of different colours. Two objects of the same shape - but of different colours - were target and distractor objects in a corresponding sentence. The video-clips each showed a sequence of camera-movements (that are called *fixations* for the human eye) towards either an object on the table or the assumed interaction partner, i.e. straight ahead. At the same time, a synthesised sentence of the following form was played back:

(1) a. "Next to the cylinder is a pyramid which is orange."
b. "Next to the <ANCHOR> is a <TARGET> which is <COLOUR>. *(as coded for analyses)*
c. "Neben dem Zylinder steht eine Pyramide die orange ist." *(original german sentence)*

The robot fixations and the spoken sentence were timed such that a fixation towards an object happened approximately one second prior to the onset of the referring noun which is consistent with psychological findings about the co-occurrence of gaze and referring expressions in human-human interaction [4, 16]. We can thus study two types of reactive human gaze: one being elicited by robot gaze (joint attention), the other being utterance-mediated (inspecting mentioned objects).

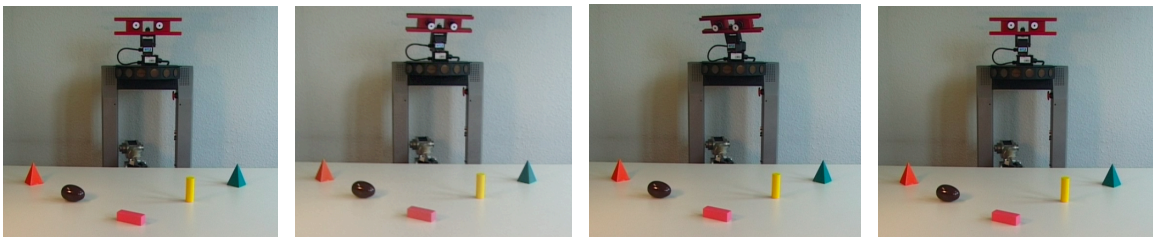


2: The timing of utterance-driven robot gaze, for sentence (1)

The presented videos were segmented into interest areas (IA) by means of bitmap templates, i.e. each video contained regions that were labelled e.g. "head", "table". Thus, the output of the eye-tracker could be mapped onto these templates yielding a certain number of hits for each IA. The spoken utterance is a sentence like example (1) describing the relation between a couple of objects. For our analysis the "cylinder" is encoded as the **anchor** reference and object, the "pyramid" is the **target** reference but may refer to the **target** object *or* the **distractor** object since there are two objects of the same shape, and the adjective "orange" is the linguistic point of disambiguation (LPoD). A similar design, also featuring late linguistic disambiguation with early visual disambiguation by means of gaze-following, was already successfully tested in a study on human-human interaction by [5].

Based on the onsets and offsets of the encoded linguistic events we segmented the video/speech stream into 7 interest periods (IP). The

[1] very kindly provided by the DFKI CoSy group: http://www.dfki.de/cosy/www/index.html and much appreciated



(a): Gaze straight ahead... (b): ...to anchor, (c): ...to distractor, (d): ...and straight ahead.

1: Frame sequence depicting a false-gaze condition (ti or fc) for sentence (1).

IPs encode the distinct time regions when the robot fixates an object and when it refers linguistically to an object (see figure 2). IP 7 is special, as it encodes the response time of the subjects, i.e. from the LPoD until the button pressing event, and is therefore a dependent variable. Because we are interested in the fixations occurring during that time period we included it as an interest period in our analyses. Note that although IP 7 varies in length it is typically longer than the other IPs and hence more fixations occur within it. This IP is to be analysed by itself with focus on the differences among the conditions.

Condition	**Spoken sentence:**
	Gaze towards:
true - congruent (tc)	Next to the cylinder is a pyramid which is turquoise.
	<cylinder> <turquoise pyramid>
true - incongruent (ti)	Next to the cylinder is a pyramid which is turquoise.
	<cylinder> <orange pyramid>
true - no gaze (tn)	Next to the cylinder is a pyramid which is turquoise.
	<no gaze>
false - congruent (fc)	Next to the cylinder is a pyramid which is orange.
	<cylinder> <orange pyramid>
false - incongruent (fi)	Next to the cylinder is a pyramid which is orange.
	<cylinder> <turquoise pyramid>
false - no gaze (fn)	Next to the cylinder is a pyramid which is orange.
	<no gaze>

3: 3 x 2 conditions and samples.

A set of six items was constructed and each item was created in all six conditions resulting in a total of 36 video clips. The six conditions are shown in figure 3. We manipulated robot gaze behaviour as follows: gaze towards the correct target in the context of the described scene, gaze towards an incorrect object and no gaze during the utterance at all. Each gaze behaviour appears with a true or false statement about the spatial relation between two objects. The result is a set of six conditions: a true statement with no gaze (tn), with congruent correct gaze (tc) or with gaze towards an incorrect distractor object (ti), and a false statement combined with no gaze (fn), congruent gaze towards the mentioned but incorrect distractor object (fc) or incongruent gaze towards the correct but not mentioned object (fi). As shown in example (1) and figure 3 the general sequence of events in each item is as follows: the robot fixates the anchor object and then refers to the anchor linguistically, then - depending on the condition - the robot looks at the target or distractor object and subsequently it refers to either object linguistically before reaching the point of linguistic disambiguation (LPoD) which is the utterance of the colour towards the end of the sentence. The robot then looks back up towards the interlocutor. Note, that in the *no gaze*-conditions, the robot performs a quick glance at the visual scene before starting to speak and then remains still. This is to ensure that even though there is no relevant robot gaze behaviour the scene looks more or less natural.

2.2.3 Procedure

An EyeLink I head-mounted eye-tracker monitored participants' eye movements. The video clips were presented on a 21-inch color monitor. Viewing was binocular, although only the dominant eye was tracked and participants' head movements were unrestricted. For each trial, a video was played and its last frame remained on the screen until an overall duration of 11 seconds was reached. After a drift correction interlude the next video clip was presented. Prior to the experiment, the participants were instructed by a short text to attend to the scene and decide whether the robot was right or wrong. They were told that the results were used as feedback in a machine learning procedure for the robot. Next, the camera was setup and calibrated manually using a nine-point fixation stimulus. The entire experiment lasted approximately 25 min.

2.2.4 Predictions

If, indeed, this cognitively motivated robot gaze behaviour is beneficial, we expect incongruent gaze behaviour to cause a slow-down in cognitive processing measurable by recording decision/response times and possibly disruptions in fixations. Concerning response times we expect generally slower response times for false statements as this is a typical effect reported of in the literature (e.g. in the case of response times for match/mismatch tasks [15]). The three gaze conditions by themselves (*congruent, incongruent, no gaze*) are also expected to yield differences: congruent gaze should facilitate understanding and elicit faster response times than incongruent gaze. In the neutral *no gaze*-conditions there are two possibilities. Either this condition elicits the fastest response times because participants generally pay little attention to the robot's gaze, simplifying on-line information complexity. Or the neutral condition's response times lie between the *congruent* and *incongruent* conditions since there is neither supportive nor disruptive information conveyed.

With respect to participants' fixations we expect to observe gaze-following. That is, we predict that people fixate those objects or regions that the robot looks to. When the robot gazes towards the (incorrect) distractor object we still predict an increase in (gaze-mediated) looks towards the distractor by the participants. Generally, we anticipate that incongruent gaze behaviour - when robot gaze and robot utterance refer to distinct and incompatible objects - will elicit saccades between these two objects (target and distractor). Furthermore, we expect to observe utterance-mediated gaze.

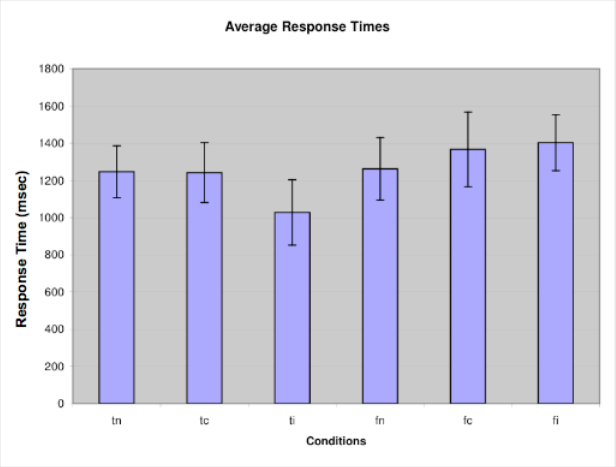
Once the robot's speech identifies an object or scene region we predict increased looks by our participants towards this region.

The most critical IPs, with regard to our predictions on gaze-mediated looks, are IP 1 and 3, which correspond to robot gaze movements. Most critical, with respect to utterance-mediated fixations, are IPs 2 and 4 (and possibly the subsequent one respectively) which correspond to the periods when the robot refers to an object linguistically. Further we similarly anticipate saccades between the target and the distractor during the response time period (IP 7) because we expect people to visually assure their decision and check all possible referents before giving the answer.

3. RESULTS

Response Times

The analysis of the response times, i.e. the time from the mentioning of the LPoD until a button press for either 'true' or 'false' was recorded, revealed that the *false-incongruent* condition (fi) results in the slowest response (figure 4). An unpaired t-test confirmed that *fi*-responses were significantly slower than *ti*-responses (*true-incongruent*): the difference in means is 375.13 msec with a 95% confidence interval of $375.13 \pm 1.96 * 117.85 = (144.14, 606.1)$ with t(*ti,fi*)$= 3.18 > t(p < 0.001, df = 128)$.



4: Average response times in msec for each condition, including upper and lower bounds for

The expected and observed general tendency for wrong statements to elicit longer response times than true statements is apparent in the graph as well. The *no gaze*-conditions are neither faster nor slower than the gaze-conditions which suggests that people do make use of robot gaze and are not finding it distracting or annoying (even though it often is wrong in this study). The slow response time for *false-incongruent* trials suggests that the participants had difficulty to determine correctness especially when a statement was false (i.e. the robot referred to the wrong object) although the robot was fixating an object that would have been correct to mention in this situation. This is consistent with our hypothesis that robot gaze is useful. In particular when it is used correctly by the robot, the gaze modality becomes a competitor to the language modality - at least in those cases where the utterance conveys unexpected or wrong information.

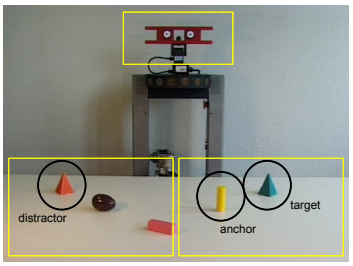
Furthermore the condition *true-incongruent* yields considerably faster response times than the other two *true* conditions (tc, tn). This initially surprising result still supports the hypothesis that gaze is useful - even when it is wrong - by yielding faster results than the *no gaze*-conditions. Considering the design of the pilot study, i.e. without fillers and the same gaze and sentence pattern for each trial, it is not surprising that people adjust and learn to recognise wrong gaze behaviour faster. In fact, this may cause some distortion of the response times in general. However, when both robot gaze and the spoken sentence are congruently referring to a wrong object (i.e. the statement is logically false), the response times are still considerably slower than the remaining four conditions (tn, fn, tc, ti). That suggests that even though both modalities are wrong, and obviously so, their congruency elicits longer response times and, hence, seems to pose a higher cognitive load.

Another interesting effect is revealed by the number of incorrect answers and those that were not given at all. It occurred several times that subjects did not press a button at all. Out of 8 omitted answers, 6 occurred in a *true-no gaze* condition and 2 in a *true-incongruent* condition. Incorrect answers were given in 22 trials, out of which 14 occurred in an *incongruent* condition and 5 in a *no gaze* condition. This makes an overall error of 7 % of all trials. The omitted and incorrect answers in trials featuring *incongruency* add up to 16 (53 % of all errors) and in trials without any directed robot gaze there are 11 incorrect answers (37 % of all errors), whereas only 3 incorrect answers where found in *congruent* trials. Again, this supports the claim that (congruent) gaze contributes to successful understanding even when produced by a robot.

Fixations

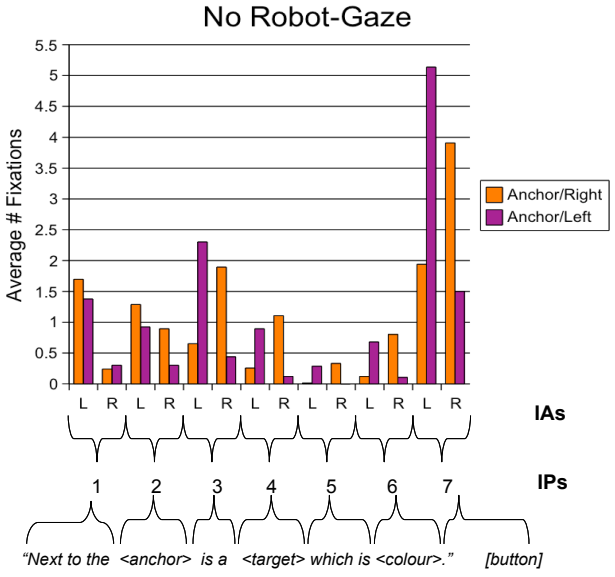
An initial analysis of the average number of fixations over all subjects per condition and per interest period (IP) and interest area (IA), here robot head and table, shows that there is a general rise in absolute number of fixations to the table area as soon as the first object is mentioned. After a slight decline towards the end of the sentence the number rises again considerably in the time period between the LPoD and the moment the subject presses the button. During the same IP the average number of fixations on the robot head rises as well. This may be due to the relief of concentration after the sentence has ended (and people had time to inspect the head) or may simply be the default gaze direction, i.e. straight ahead.

In a more fine-grained analysis, we have looked at three IAs, one being again the robot head and two more where the table area has been divided into two parts, left and right. In half of the trials the anchor object and the target object are positioned in the right half of the table area whereas the distractor object lies in the left area of the table, and vice versa for the other half. We therefore refer to the area that contains the referent and target objects as the target area and to the other side of the table as the distractor area. This kind of segmentation allowed us to observe whether participants followed the robot's gaze movement without fixating the



5: IAs 'head', 'left' and 'right'

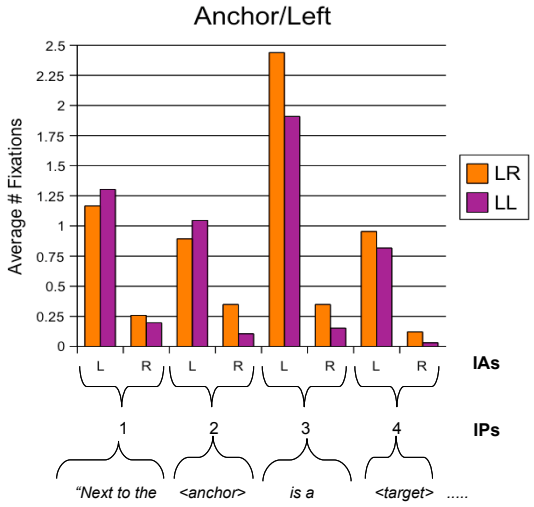
robot head directly. As a result we observed a general bias for fixations on the left side of the table. This becomes evident in figure 6 which shows the control condition *no gaze* (tn, fn). IP 1 and 2 in this chart show a clear preference for looks towards the left side of the table independently of where the anchor/target objects are positioned. This bias is commonly observed, reflecting general human scan patterns. From IP 3 on, however, utterance-driven fixations can be observed showing the expected preference for the area containing the mentioned anchor/target.



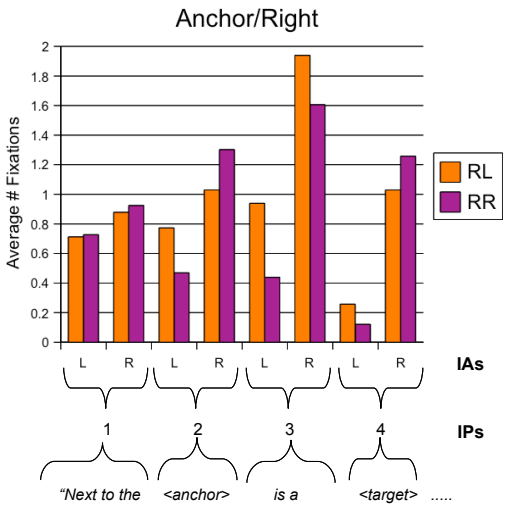
6: Average number of fixations in the neutral *no gaze*-conditions, with target being on either the right or left side. L and R on the x-axis denote the IAs 'left' and 'right'.

In figure 7 we have plotted the average number of fixations in IPs 1-4: 1, 3 depicting mainly robot gaze-mediated fixations (joint-attention) and IP 2, 4 showing mainly utterance-mediated fixations. Note, that RR, RL etc. denote the direction and therefore congruency of gaze movements, i.e. towards the right side of the table (target area) and again rightwards is correct gaze movement, the abbreviation is thus RR. RL indicates gaze movement towards the anchor (target area) and then to the left side of the table where the distractor object is located. For IPs 1 and 3 we detected more fixations on the anchor/target area than in the distractor area, notably already before the object was actually mentioned. The result is significant according to the paired t-test, for instance in IP 3 for the target on the left side: mean difference $\bar{x}_L = 1.92$ fixations in a confidence interval (1.68, 2.15) with $t(\bar{x}_L) = 16.13 > t(p < 0.001, df = 131)$. And accordingly for the target on the right side: mean difference $x_R = 1.08$ fixations in a confidence interval (0.805, 1.354) with $t(\bar{x}_R) = 7.714 > t(p < 0.001, df = 131)$.

The same effect, i.e. a significant rise in fixations towards the object that is (now linguistically) referred to, is visible in IPs 2 and 4 and is continued throughout the rest of the trial. The most critical region is IP 3 where the robot's gaze is either turned towards the target or the distractor object at the other side of the table. At this stage it becomes evident whether subjects believe that the robot gaze is an early indicator of what is going to be mentioned next. Our recordings reveal only a slight increase of fixations towards the distractor as a reaction to robot gaze towards the distractor object. The experimental design may prevent a stronger effect because the repeated gaze pattern in the items allows the participants to predict what is going to happen.



(a): Here IA "L" contains anchor and target...

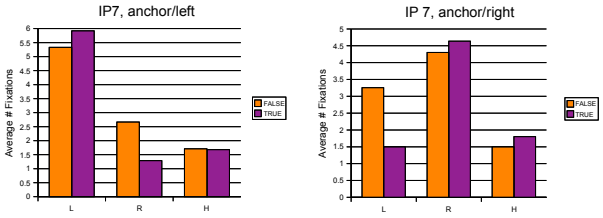


(b): ...while here IA "R" contains anchor and target.

7: Anchor/target on one side of the table, distractor on the opposite side. Depicted are IPs 1-4 showing gaze- and utterance-mediated fixations.

IP 7 is plotted in figure 8 which depicts the distinct conditions within the decision time period, i.e. from the disambiguating adjective until the button press event. This reveals that a false statement leads to a significantly higher number of fixations in the distractor area than is the case for true statement trials: mean difference of average fixations on the distractor area and on target area (for target/left side) is $\bar{x}_L = 1.38$ fixations in a 95% confidence interval (0.645, 2.115) with (un-paired) $t(\bar{x}_L) = 3.373 > t(p < 0.001, df = 130)$, and similarly for target/right side: $\bar{x}_R = 1.755$ fixations in a confidence interval (1.003, 2.507) with $t(\bar{x}_R) = 4.5726 > t(p < 0.001, df = 130)$. This result is not surprising as giving wrong answers typically affords slightly longer response times (as reported in the previous section) which should also be reflected in the direction and number of fixations performed during

that time. However, we did not observe the expected difference with respect to congruency at this stage. That is, a true statement yielded similar fixation results independently of the correctness of the robot-gaze. It is likely that this is due to the relatively easy spatial arrangement and the long time period until a decision is demanded such that participants can look around extensively beforehand.



(a): Distractor object lies on the right...

(b): ...and here on the left side.

8: Fixations for true/false statements during response time period.

4. DISCUSSION

We found clear evidence for (robot) utterance-mediated gaze in human-robot interaction: people look to the objects that the robot refers to linguistically. This is not a surprising result but it is useful nonetheless as it confirms typical human behaviour in response to robot speech and gaze (even in video-based interaction). Further evidence was collected during the response period: we registered a strong tendency of the participants to fixate both the target and the distractor object when the statement was false. In those cases, the uttered sentence referred to the distractor object linguistically and people looked more often towards the distractor object than in those trials where a true sentence was uttered. We also found clear evidence for gaze-mediated joint attention, i.e. people look to objects that the robot looks at. IP 1 was the period immediately preceding the linguistic reference to the anchor and already then participants looked towards the anchor. These results support our hypothesis that human-robot spoken interaction is governed by principles similar to HHI. When the robot's language-related gaze behaviour is similar to that of people we observe human gaze patterns that are typical for HHI.

Moreover, the reported response accuracies suggest that generally incongruent robot behaviour (i.e. divergence of both modalities speech and gaze) is causing confusion. The measured response times are slightly more difficult to interpret. False statements elicited slightly longer response times which is typical human behaviour. We also found that the *false-congruent* condition response times were significantly slower than in the *true-incongruent* condition (which seems to contradict the evidence from response accuracy). However, this could suggest that the coherence of the modalities in the (wrong) fc-condition leads to stronger doubts about the truth of the statement than in the ti-condition. The ti-condition in which robot gaze is wrong while the linguistic statement is true seems to allow fast reference resolution. Considering the design of the pilot study, i.e. without fillers and the same repetitive gaze and sentence pattern in each trial, we assume that people adapt to the task and learn to recognise early when gaze is erroneous which may generally distort response times.

The effects we found were not always in accordance with our predictions. We assumed, for instance, that incongruent robot-behaviour elicits more fixations on both potential referents, target and distractor, during the linguistic utterance and during decision making. This was partially observed in the latter interest period for false statements. The direction of the robot gaze, however, seemed to be irrelevant for the final decision process. For wrong gaze we did not observe a particular rise in fixations towards the distractor area in IP 3 either. This IP comprises the robot-gaze movement towards the target or distractor object and, thus, reports gaze-following. Presumably this again is due to the fact that the course of events in a trial becomes predictable after a while.

5. CONCLUSIONS AND FUTURE WORK

We have shown that, in principle, it is possible to use detailed insights into human cognition and behaviour to enrich human-robot-interaction. The presented evidence shows that this kind of robot-gaze is beneficial in HRI and that humans react in a manner typical of HHI to both robot speech and robot gaze. We predicted that in case one or both robot modalities are infelicitous a slow-down of the interaction would be measurable by response times and fixation distributions. Our results support this hypothesis and reveal several cases where incongruent robot behaviour leads to slower response times or disruptions in the usual distribution of fixations.

The presented study also shows that the methods we used to measure the effects and success of the robot behaviour objectively during robot production are generally appropriate and effective. However, we found some weaknesses in the experimental design such that the obtained results, although promising, are only preliminary. The spatial arrangement of the scene, for instance, is small and simple. A larger area with more complexity could lead to clearer results concerning the robot gaze-mediated fixations, i.e. the robot gaze could be considered more useful for early referent resolution. The presentation of the items is going to be interleaved with the presentation of filler videos which differ from the items and enforce gaze reliability. This ensures that the participants will not be able to predict what the robot is going to do. More crucially, the presentation of mostly *true-congruent* fillers will influence the trade-off between cost and benefit of robot-gaze for information processing during communication: the more errors occur in the trials the likelier will participants decide to ignore gaze as a source of information. If, however, gaze is mostly a useful early indicator for correct reference resolution, its benefit for on-line comprehension should override the extra costs caused by errors. Furthermore, we have put a lot of emphasis on congruency in this study but designed a task for the subjects that focusses on the truth value of the linguistic statement alone. In order to emphasise the effect of incongruency we plan to change the task such that subjects are required to consider the robot performance more holistically, e.g. "If you think the robot is wrong, tell it what is wrong". There is a trade-off here between simplicity in measurements: we lose the option to record response times but we gain additional information on what the participants expected of the robot and what they actually perceived. This is possible because incongruency arises from different errors made by the robot: either the linguistic reference is incorrect or the gaze is incorrect. This failure may be attributed to incorrect wording (as in 'correct gaze but false statement' = ti), incorrect judgement of the spatial relations, errors in visual (colour) processing, erroneous gaze movement etc. Thus, by asking the participants to actively decide which object was meant and why the error occurred, we also learn more about the "theory-of-mind" (ToM) that people ascribe to the robot and how robot gaze contributes to forming a "theory-of-mind" (see e.g. [1] for work on ToM among humans and [11] for

research on apes). With increasing communicative competence of robots it becomes more and more interesting to investigate people's attitude towards robots.

6. ACKNOWLEDGMENTS

The research reported of in this paper was supported by IRTG 715 "Language Technology and Cognitive Systems" funded by the German Research Foundation (DFG).

7. REFERENCES

[1] S. Baron-Cohen, A. Leslie, and U. Frith. Does the autistic child have a "theory of mind"? *Cognition*, 21:37Ð46, 1985.

[2] J. Cassell, O. Torres, and S. Prevost. Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation. *Machine Conversations*, pages 143–154, 1999.

[3] P. D. Chris Moore, Philip J. Dunham, editor. *Joint Attention Its Origins and Role in Development*. Lawrence Erlbaum Associates, 1995.

[4] Z. M. Griffin and K. Bock. What the eyes say about speaking. *Psychological Science*, 11:274–279, 2000.

[5] J. Hanna and S. Brennan. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57:596–615, 2007.

[6] M. Imai, T. Kanda, T. Ono, H. Ishiguro, and K. Mase. Robot mediated round table: Analysis of the effect of robot's gaze. In *Proceedings of 11th IEEE ROMAN '02*, pages 411–416, 2002.

[7] P. Knoeferle and M. Crocker. The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*, 30:481–529, 2006.

[8] P. Knoeferle and M. Crocker. The influence of recent scene events on spoken comprehension: evidence from eye-movements. *Journal of Memory and Language (Special issue: Language-Vision Interaction)*, 57:519–543, 2007.

[9] A. Meyer, A. Sleiderink, and W. Levelt. Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, 66:B25–B33, 1998.

[10] A. Meyer, F. van der Meulen, and A. Brooks. Eye movements during speech planning: Talking about present and remembered objects. *Visual Cognition*, 11:553–576, 2004.

[11] B. H. Michael Tomasello, Josep Call. Chimpanzees versus humans: it's not that simple. *Trends in Cognitive Sciences*, 7:1632–1634, 239-240.

[12] B. Mutlu, J. Hodgins, and J. Forlizzi. A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior. In *Proceedings 2006 IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS'06)*, Genova, Italy, 2006.

[13] C. L. Sidner, C. Lee, C. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164, 2005.

[14] M. K. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634, 1995.

[15] G. Underwood, L. Jebbett, and K. Roberts. Inspecting pictures for information to verify a sentence: eye movements in general encoding and in focused search. *The Quarterly Journal of Experimental Psychology*, 56:165–182, 2004.

[16] F. F. van der Meulen, A. S. Meyer, and W. J. M. Levelt. Eye movements during the production of nouns and pronouns. *Memory & Cognition*, 29(3):512–521, 2001.

[17] S. Woods, M. Walters, K. L. Koay, and K. Dautenhahn. Comparing Human Robot Interaction Scenarios Using Live and Video Based Methods: Towards a Novel Methodological Approach. In *Proc. AMC'06, The 9th International Workshop on Advanced Motion Control*, 2006.