The Dynamics of Referential Speaker Gaze: Order is Important, Synchronization, Not So Much

Maria Staudte & Matthew W. Crocker
Department of Computational Linguistics
Saarland University
Saarbrücken, Germany
{masta, crocker}@coli.uni-saarland.de

1 Introduction

Gaze is known to be an important social cue in face-to-face communication indicating interest, focus of attention, emotional states, or intentions to give and take speaking turns [3]. In addition, speaker gaze can also influence object perception [2] and situated utterance comprehension [6] by driving both interlocutors visual attention towards the same object. Such gaze-following can lead to "joint attention" [3, 4] to an object, helping to ground and disambiguate referring expressions [6] and to coordinate actions in pursuing a shared goal [11]. The precise temporal and causal processes involved in on-line gaze-following during concurrent utterance comprehension are, however, still largely unknown. Specifically, does speaker gaze really benefit listeners? And if so, what temporal relationship of gaze and speech is essential to such benefit? In this paper, we review several of our recent findings [8, 9] that have exploited artificial agents (both robots and virtual agents) to systematically investigate how speaker gaze influences listeners on-line comprehension and to shed light onto the dynamics of integrating real-time multi-modal information.

The use of artificial agents as speakers circumvents variation in speaker gaze, enabling greater uniformity and control of the gaze-speech synchronization, while the fallibility of such agents also makes the investigation of non-natural gaze behavior more plausible. Previous research has shown that artificial agents are valuable tools for investigating social behaviors [7] which cause reactions that are similar to those observed in human-human interaction [6]. In our experimental setting, an artificial agent uttered statements about objects in the visual context, while listeners were told to determine the validity of the agents statement as quickly as possible. Listener eye-movements and task performance were both analysed. In a series of studies, this setting was used to investigate: (a) the dynamics of concurrently processing speaker gaze (both gaze movement and fixation) and utterance (Experiment 1 [party published in 8]); (b) to determine the importance of temporal synchronization and linear order alignment between gaze cues and the corresponding referring expressions (Experiment 2 [9]); and (c) to determine the role of linear order in greater depth using a different agent (Experiment 3). Taken together, our findings point to the clear benefit of speaker gaze for listeners, but suggest that while the linear order of gaze and object mention must be maintained, close temporal synchronization of gaze and speech is less critical.

2 Experiment 1: Gaze Benefit

While listeners are known to rapidly fixate objects mentioned by a speaker [10], it is also possible that they will exploit speaker gaze to these object in advance of their mention [5]. In this experiment, we used a robot to examine whether referential speaker gaze, that is aligned to speech in a human-like manner [5], is followed and how such gaze along with simultaneously perceived referring expressions drive listeners’ visual attention [8]. Participants watched a videotaped robot produce statements about several objects in its view, and were asked to judge its utterances for validity with respect to the shared scene.
A description such as "The cylinder is taller than the pyramid that is pink." was further accompanied by robot gaze movements – first to the cylinder, then either to the small pink pyramid (congruent), a tall brown pyramid (incongruent), or to no object at all (neutral) – each occurring shortly before the robot uttered the corresponding noun phrases [5]. To determine whether listeners followed robot gaze to a mentioned referent, the second noun phrase was temporarily ambiguous with respect to its referent in the scene. That is, the scene provided two pyramids of different sizes and colors, one of which the robot mentioned using the according color adjective (Fig. 2). The temporal ambiguity provided time for participants to integrate the gaze cue with the ambiguous noun before the color was even mentioned. This 2x3 manipulation (Statement Validity: true/false; Gaze Congruency: congruent, incongruent, neutral) manipulation resulted in six conditions. Two dependent measures enabled us to determined the objective benefit of speaker gaze (in the judgement task reaction times), while also teasing apart the influence of gaze/head movement, final gaze direction, and the unfolding linguistic material as revealed by listener eye movements.

Materials & Procedure: All conditions of an item video showed the same scene and only differed with respect to where the robot looked and whether the utterance was correct (i.e., the robot verbally referred to the correct (target) object or the incorrect (competitor) object). Before the experiment, participants were asked to attend to the presented videos and judge whether or not the robot’s statements were valid with respect to each scene. Crucially, speaker gaze was typically not required nor did it change the assessment of sentence validity with respect to the scene. An EyeLink II head-mounted eye-tracker monitored participants’ eye movements, and their button press latencies were recorded to assess the time needed to comprehend and validate an utterance.

All objects in the scene as well as the robot head were coded as Interest Areas and, based on the speech stream, the video was segmented into the following adjacent Interest Periods: the gaze movement (IP0), the fixation period (IP1), the utterance of the referring noun (IP2), and finally the utterance of the disambiguating adjective (IP3). The analyses were carried out using mixed-effect models from the lme4 package in R and Chi-Square tests to assess the contribution of a predictor through model reduction [1].

Results: To exemplify listener behavior, Fig. 2 shows a plot of the eye-movement data for one condition – a true sentence with incongruent gaze (towards the brown competitor pyramid) – for the duration of a whole trial. Initially, people looked mainly at the robot head. Only when the robot head moved towards the cylinder and, more clearly, when the robot started speaking, listeners directed visual attention away from the robot’s head and towards the cylinder. Throughout the course of a trial, listeners rarely looked back at the robot head. The plot as well as inferential statistics, however, clearly indicate gaze-following to the cylinder and the corresponding pyramid (the brown competitor pyramid in the plot of true-incongruent trials) which suggests that robot gaze was used peripherally. Moreover, analyses of looks towards the distractors located between the cylinder and the two pyramids reveal a sweeping gaze behavior of listeners, i.e., they seem to follow what is within the scope of the robot’s gaze movement. Again, Fig. 2 illustrates how listeners initially look at the distractor (small egg, in IP0) before increasing looks into the final direction of robot gaze (at the brown competitor pyramid, in IP1).

During IP3 ("pink/brown"), listeners significantly more frequently looked at the object fixated by the robot, until the color adjective either confirms or rejects this object as actual referent. Consequently, main effects were also found for both Sentence Validity ($\chi^2(1) = 19.06; p < 0.001$) and, crucially, Gaze Congruency ($\chi^2(2) = 60.43; p < 0.001$).
in the response time data. That is, participants responses were faster when gaze and speech were congruent, and slower when they were incongruent, compared to the neutral case. Overall, these results suggest that gaze is a dynamic cue, consisting of a movement phase (comparable to human head movement) and a final fixation phase, which are both used and integrated with the unfolding utterance.

3 Experiment 2: Temporal Alignment and Gaze Order

A second study further explored the role of gaze and speech alignment. Specifically, we investigated whether referential speaker gaze needs to be temporally synchronized with speech (in the way human gaze is typically synchronized) in order to be beneficial, or whether speaker gaze conveys referential intentions that have a more persistent effect on utterance comprehension. In this study, robot gaze was always directed to the mentioned objects, and we manipulated the linear Order of gaze cues, as well as their Synchronization with the utterance.

Materials & Procedure: Sentences were similar as in Experiment 1 but contained no temporary ambiguity, e.g. "The cylinder is taller than the pink pyramid." The manipulation of Order (linearly aligned, reverse) and Synchronization (synchronized, preceding) yielded four conditions for each item. The temporal delay between gaze and speech was roughly 5.3 seconds in the preceding condition – i.e. the robot first looked at cylinder and pyramid and then began the utterance – versus a 1 second delay in the synchronized condition, as in Experiment 1. In condition linearly aligned & synchronized, gaze and speech cues were coherent and synchronized in a human-like manner while the condition reverse & synchronized showed cues that were concurrent but reverse to each other.

Results: Again, gaze-following was observed in all conditions as well as referential looks elicited by the uttered referring expressions. Interestingly, (temporal) Synchronization had no effect on listeners’ response times. That is, participants were equally fast to determine the validity of the robot statement in synchronized and preceding conditions. Model reduction further revealed a main effect of Order ($\chi^2(1) = 45.19, p < .001$). This effect of Order suggests that reverse order was indeed more difficult to process than linearly correctly ordered cues, probably due to the increased effort in resolving the concurrent and conflicting referential information.

The presented results thus suggest that the order of gaze cues must be similar to order of mention. Large temporal shifts of robot gaze with respect to its ‘natural’ synchronization, however, do not substantially affect the utility of the gaze cues. We conclude, therefore, that the utility of speaker gaze results from it revealing the speakers referential intentions, rather than serving simply as a synchronized visual cue to the mentioned object.

4 Experiment 3: Facilitation and Disruption due to Gaze Order

Our final study sought to replicate the findings regarding gaze order using a virtual agent with more human-like eye and head movement and slightly more complex scene, while also adding a neutral gaze condition. The motivation for this was to establish a baseline to quantify the negative effect of reversely ordered cues on comprehension. Materials and procedure were similar in Experiments 2 and 3, but crucially, the videos now showed a virtual agent that described objects in it’s virtual scene and that produced faster and more human-like head and eye movements. Analyses of the eye movement data confirmed previously observed gaze and speech following behavior. In the reverse condition illustrated in Fig. 3, this led to fixations first to the pyramid (second noun) and only then to the mentioned cylinder (first

Figure 3: Fixations in reverse order trials.
This “wrong” order of looking at objects with respect to their mentioning seemed to disrupt listeners’ comprehension compared to neutral or correctly aligned gaze: Response time data revealed a main effect of Order (linearly aligned: 1340.9ms, neutral: 1809.9ms, reverse 2062.9ms, $\chi^2(2) = 68.82, p < .001$). Interestingly, the point when people shift attention to the first mentioned object, i.e., the cylinder, is identical in the neutral and the reverse condition and, yet, response times in both conditions differ significantly. It seems that, even though listeners are able to resolve linguistic references in the same time, the fact that attention was initially drawn to the pyramid interferes with the actual utterance.

5 Conclusion

Results from Experiment 1 suggest that gaze is a dynamic cue, consisting of a movement phase and a final fixation phase, which are both used and integrated with the unfolding utterance. This supports the view that gaze and speech are used simultaneously to incrementally update a common, multimodal representation of what is assumed to be the proposition intended by the speaker. Experiments 2 further reveals that large temporal shifts of robot gaze with respect to its ‘natural’ synchronization do not substantially affect the utility of the gaze cues whereas the order of cues does. If this expectation is not met, gaze cues may even disrupt comprehension, as the comparison with neutral gaze suggests in Experiment 3. Thus, people appear to expect that the inferred referential intentions be realized in the corresponding order of speaker fixations[5], but are robust to variation in their synchronization. We believe that this result is particularly interesting for the development of autonomous gaze and speech generation in artificial agents as it reveals which behavioral features are important for efficient communication (e.g., linear order of referential cues) and which may be neglected during production (e.g., temporal synchronization). Modeling these dynamics of order in attention shifts (as induced by speaker gaze) along with incremental utterance comprehension may potentially contribute to understanding when and how visual attention to potential referents influences the internal (re-)construction of a proposition.

References