On the importance of gaze and speech alignment for efficient communication

Maria Staudte[1], Alexis Heloir[2], Matthew Crocker[1] and Michael Kipp[2]

[1]Department of Computational Linguistics, Saarland University, Germany
{masta,crocker}@coli.uni-saarland.de

[2]DFKI, Embodied Agents Research Group, Germany
firstname.surname@dfki.de

Fig. 1. Sequence of congruent multimodal references as performed by the agent.

Keywords: referential gaze, spoken interaction, virtual speaker, alignment

Gaze as Visual Reference

Gaze is known to be an important social cue in face-to-face communication indicating focus of attention. Speaker gaze can influence object perception and situated utterance comprehension by driving both interlocutors' visual attention towards the same object; hence facilitating grounding and disambiguation [1]. The precise temporal and causal processes involved in on-line gaze-following during concurrent utterance comprehension are, however, still largely unknown. Specifically, the alignment of referential gaze and speech cues may be essential to such benefit. In this paper, we report findings from an eye-tracking study exploiting a virtual character [2] to systematically assess how speaker gaze influences listeners' on-line comprehension.

Firstly, we provide supporting evidence for the hypothesis that artificial characters in general, and our character in particular, can serve as a valuable tool to study how listeners integrate real-time gaze and speech. Secondly, our findings point to a clear benefit of speaker gaze for listeners when gaze cues and verbal references occur in identical order. In contrast, inconguent gaze is shown to have a disruptive effect on comprehension.

Experiment

24 Participants watched videos of the virtual character producing statements about several objects in her view, and were asked to judge these utterances for validity with respect to the scene (Fig. 1). We manipulated the sequential order of gaze gestures to mentioned objects and spoken referring expressions in order to investigate the importance of the sequential congruency between gaze and speech. That is, a description such as ”The star is taller than the pyramid.” was accompanied by head and eye movements – either first to the star, then to the pyramid (congruent, see Fig.1), or first to the pyramid, then to the star (incongruent), or to no object at all (neutral). Each gaze movement occurred shortly before the character uttered the corresponding noun phrases.

The agent used in this experiment was controlled by the EMBR framework [3] using the BML language [4]. The different channels involved in the multimodal behaviors and specified in BML can be synchronized using symbolic tags which are resolved in the realization phase. The BML sequence we actually used to produce the depicted behavior is provided below. Lexemes ”R1L2A”, ”R1L2B”, and ”R1L2C”, represent links to the respective gaze gestures.

```
<bml>
        <speech id="s1">
                <text> <sync id="1"/> The star is taller than <sync id="2"/>
                        the pyramid <sync id="3"/>
                </text>
        </speech>
        <gesture  id="g1"  type="LEXICALIZED"  lexeme="R1L2A"  end="s1:1"  />
        <gesture  id="g2"  type="LEXICALIZED"  lexeme="R1L2B"  end="s1:2"  />
        <gesture  id="g3"  type="LEXICALIZED"  lexeme="R1L2C"  start="s1:3"  />
</bml>
```

The EMBOTS framework uses the Open Mary speech synthesizer [5] and the EMBR character animation engine [2] to produce the spoken utterances and the animated gestures, respectively, that were used in this experiment.

Results

We used two independent measures to assess the objective benefit of speaker gaze: Eye-movements during comprehension and response times in the judgment task. The eye-movement data is plotted in Fig. 2 and the results confirmed previously observed gaze and speech following behavior [6]. Specifically, we found that, in the congruent condition, gaze-following resulted in listener fixations to the star and the pyramid even before their mention. In the reverse condition, however, gaze-following guided listeners’ visual attention towards the pyramid before they heard the first noun ”star”. Then the noun as well as the agent’s next gaze movement drew listeners’ visual attention to the star - before finally hearing the second noun ”pyramid”.

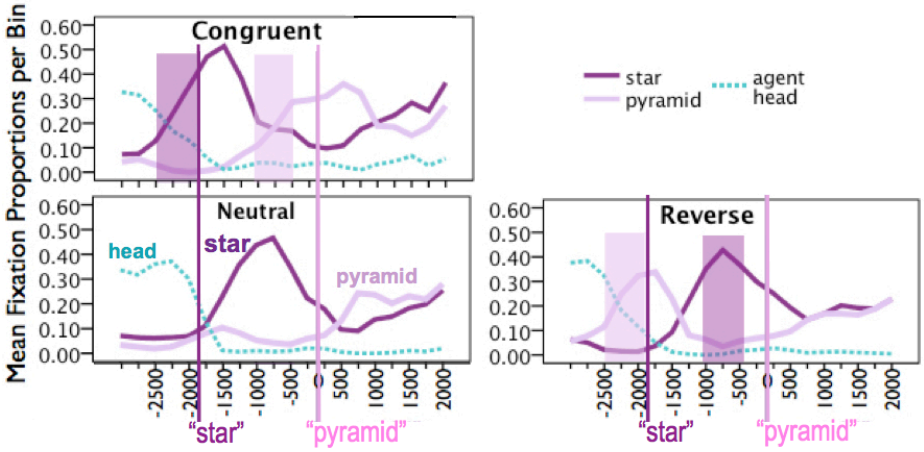

Fig. 2. Fixations across an unfolding sentence in trials, in all three conditions.

Further, the mean response times are depicted in Figure 3. An ANOVA, run on the linear mixed-effects model fitting the data, revealed a main effect of cue order ($F = 53.42$, $df = 2$). These results suggest that, in the congruent condition, gaze-following did indeed lead to an anticipation of the object mentioned next which facilitated validation, compared to the neutral case. In the incongruent condition, however, listeners' comprehension was disrupted, as revealed by the slowed response time.

Interestingly, the eye-movement data suggests that even though listeners resolved linguistic references at the same time as in the neutral case, the fact that their attention was initially drawn to another object slowed validation. This may indicate that the agent's gaze did not only increase the salience of an object but that it further elicited a concrete expectation for this object to be mentioned next. When this expectation was violated (in the reverse condition), it apparently complicated the integration process of the visual and linguistic information.

Discussion

We created a sequential mismatch between visual and spoken references by manipulating cue order, enabling us to observe which reference participants followed initially and how they recovered from a mismatch. Considering the regularity of the fixation patterns when agent gaze was available, both orders potentially offer advantages over the neutral agent gaze condition: While congruent gaze enables visual anticipation of each next referent, the reverse gaze condition offers an initial visual indicator to the second referent (the pyramid). Along with the first spoken reference to the ”star”, these two pieces of information in fact reveal both referents even earlier during the utterance than in the congruent or neutral conditions. Nevertheless, the response time data suggests that listeners are disrupted by a reversed cue order, which highlights the importance of gaze and its relative ordering for reference resolution.

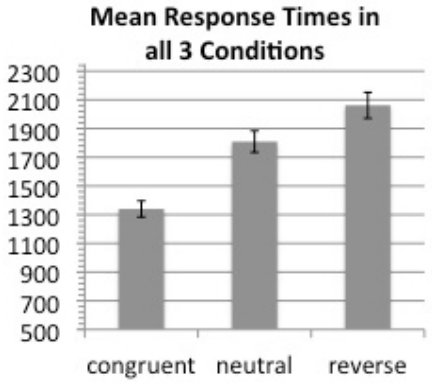

Fig. 3. Avg. response times in all three conditions.

Thus, we conclude that, firstly, character gaze drives people's attention in much the same way that human gaze does, which supports the validity of such an experimental design for general investigations of gaze and speech processing. Secondly, referential gaze and speech cues need to occur in a congruent sequence to have beneficial effects on utterance comprehension. This finding not only provides insights on how to craft natural and efficient gaze behavior for embodied agents but also sheds light on how humans interpret referential gaze.

References

1. Hanna, J., Brennan, S.: Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language,* 57(4), 596–615, (2007).

2. Heloir, A., Kipp, M.: Realtime animation of interactive agents: Specification and realization. *Journal of Applied Artificial Intelligence,* 24(6), 510–529. Taylor & Francis, (2010).

3. Kipp, M., Heloir, A., Gebhard, P., Schröder, M.: Realizing multimodal behavior: Closing the gap between behavior planning and embodied agent presentation. In: *Proceedings of the 10th Int. Conf. on Intelligent Virtual Agents*, Springer, (2010).

4. Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjlmsson, H.: Towards a common framework for multimodal generation: The behavior markup language. In: *Proceedings of the 6th Int. Conf. on Intelligent Virtual Agents*, Springer, (2006).

5. Schröder, M., Hunecke, A., Krstulovic, S.: OpenMary - open source unit selection as the basis for research on expressive synthesis. In: *Proceedings of the Blizzard Challenge* 2006.

6. Staudte, M., Crocker, M.W.: When Robot Gaze Helps Human Listeners: Attentional versus Intentional Account. In: Ohlsson, S., Catrambone, R. (eds.), *Proceedings of the 32nd Annual Conf. of the Cognitive Science Society*. Portland, OR: Cognitive Science Society, (2010).