Eye gaze cues for coordination in collaborative tasks

Susan E. Brennan
Departments of Psychology &
Computer Science
Stony Brook University
Stony Brook, NY 11794 USA
susan.brennan@sunysb.edu

Gregory J. Zelinsky
Departments of Psychology &
Computer Science
Stony Brook University
Stony Brook, NY 11794 USA
gregory.zelinsky@stonybrook.edu

Joy E. Hanna
Department of Psychology
Oberlin College
Oberlin, OH 44074 USA
Joy.Hanna@oberlin.edu

Kelly J. Savietta
Department of Psychology
Oberlin College
Oberlin, OH 44074 USA

DUET Workshop, CSCW'12, February 11-15, 2012, Seattle, Washington, USA.
ACM 978-1-4503-0556-3/12/02.

Abstract

Cues about a partner's focus of attention can be used for distributing effort and coordinating joint attention in a collaborative task. In this paper we develop some theoretical foundations and describe a research agenda for the two-way sharing of eye gaze cues in collaborative interaction. We describe our previous results on the use of gaze cues by interacting partners, both face-to-face (Hanna & Brennan, 2007; Hanna, Brennan, & Savietta, 2011) and remotely with partners seeing one another's gaze cursors in real time superimposed over a display (Brennan et al., 2008; Neider et al., 2010). We are examining the costs and benefits of partners sharing eye gaze in referential communication, collaborative search, reaching consensus, and goal recognition or "mind reading".

Keywords

Eye gaze, dual eye tracking, collaborative cognition, multimodal communication, referential communication

ACM Classification Keywords

H.5.3 [Information interfaces and presentation] computer-mediated communication, CSCW

General Terms

Human Factions, Experimentation

Introduction

In any joint human activity, such as a conversation or a collaborative task, there is a need for people to coordinate their behavior and individual contributions with one another. Such coordination depends on attending to the activity underway, recognizing what a partner is attending to, inferring the partner's intention, monitoring task progress over time, responding contingently (and at just the right moment), and, when relevant, achieving a joint focus of attention. There are abundant methods for doing this face-to-face; people can make eye contact, follow each other's gaze, monitor what their partner is oriented toward and doing (using eye gaze, head orientation, or other cues), attract the partner's attention by speaking or moving, use language, gesture, and other means to refer to and highlight relevant aspects of the task at hand, explicitly demonstrate their own understanding or describe their own contributions, and seek evidence of understanding, uptake, and task progress from the partner. These methods are easy to use when partners can assume that they have common ground by virtue of physical copresence (Clark & Marshall, 1981); being able to see where a partner is looking, and to assume that the partner can monitor one's own gaze, is especially relevant to collaboration on a spatial task. Now that eye gaze information can be transmitted between remotely located partners, the possibilities for coordinating joint attention are considerably expanded.

Grounding, multimodal cues, and the distribution of effort

We have used the *grounding* theoretical framework (Clark & Brennan, 1991; Brennan & Hulteen, 1995) to conceptualize both face-to-face and remote collaboration as coordinated activity. In short, it is not enough to simply present an utterance or send a message or perform an action; there must be evidence that a partner has perceived, understood, and repaired or taken up this information (Clark & Schaefer, 1989). The grounding framework predicts that methods for coordinating processing and behavior with a partner have the potential not only to yield benefits, but also to incur costs. Processing and behavioral resources may need to be deployed for grounding (e.g., to manage the interaction and integrate its joint product). For this reason, joint activity should not be considered simply as a summing-up of autonomous actions.

An unfolding interaction (whether during a dialogue or other collaborative activity) is incrementally shaped by the methods for coordination that are afforded by the communication medium. Methods for grounding draw on cues that can be verbal, nonverbal or both, and these cues can be redundant, augment one another, or substitute for one another. A cue may differ in how easy it is to provide or use; for example, speaking is easier than typing for most people, so spoken conversations tend to be wordier than typed ones. But spoken conversation requires partners to be auditorily copresent at the same time, and when this is difficult or undesirable, texting may be easier. The roles, capabilities, and other circumstances of two partners may differ as well. Because it may be easier for one partner to provide a particular cue than for the other to seek it out, this tends to shape who takes responsibility at a given point. For example, in one experiment, a person following the spoken directions of another person in a spatial navigation task produced more spoken responses and backchannels when she (the follower) knew that the director could not see what she was doing, but withheld such evidence when he could (essentially substituting visual evidence for a spoken "turn" in the conversation; Brennan, 2004). In this way, partners flexibly shift responsibility for who determines when one person has understood one another well enough for current purposes so that they can move on (see Clark & Wilkes-Gibbs', 1986, description of the principles of mutual responsibility and least collaborative effort).

The information available in eye gaze

Cues for coordination are not static, but unfold over time. Patterns of looking or "gaze signatures" are a promising topic for further study, as these may well be informative for intention recognition or "mind reading" between human partners, in human-computer interaction, or in tasks where cognition or behavior is augmented by technology.

Eye tracking has never been a popular input method of human-computer interface designers, as it can be unreliable (e.g., during blinking or when involuntarily captured by something in the visual environment) as well as tiring to control (see Jacob, 1995 for further discussion). It is also inherently ambiguous, serving multiple functions at once such as searching an environment, noticing coincidences or other information that may or may not be relevant, and fixating objects deemed relevant to a decision or goal. As a result, it has been used as an input modality mainly by people without other options. However, it still makes sense to consider spontaneously produced eye gaze in terms of its potential as a communicative signal, especially when deployed more or less naturally in the services of a visuo-spatial task. It may be useful to consider gaze as a kind of gesture, as typologies of gestures (e.g., McNeill, 1992) are potentially relevant to gaze. Gestures can be iconic or representational or mimetic (for eye gaze, this would correspond to a recognizable task-relevant gaze signature). Gestures can be emblems (as in conventional, culturally-specific gestures that can replace words; an example of this in the gaze domain would be rolling the eyes to convey exasperation or sarcasm). Gestures can be deictic (just as a young child points longer at an object while looking back to monitor a parent's attentional focus on the object, one may gaze longer at an object to signal its significance, essentially pointing with the eyes). And gestures can be interactive (as in open-palm gestures and "beat" gestures used to rhythmically punctuate speech; the analogue for gaze may be the regular pattern of eye contact during conversation, where addressees spend more time gazing at speakers than speakers gaze at addressees, with speakers gazing upward occasionally to display thinking about or recollecting something relevant).

Experiments: Using eye gaze cues face-to-face

As cognitive psychologists, we first used head-mounted eye gaze in our research as a dependent variable to measure such processes as language understanding (Brennan, Hanna) and visual search (Zelinsky). We began to consider eye gaze as a communicative cue that could also shape collaborative cognition in these two domains. In a program of laboratory experiments measuring the use of eye gaze by face-to-face partners referring to objects (see Fig. 1), we first found that addressees could use speakers' direction of gaze as an early cue to resolve ambiguity between same-color target and competitor objects even before hearing the disambiguating linguistic information in the unfolding referring expression. In fact, even when pairs sitting across a table from each other knew their displays were reversed (what was to the speaker's left was to the addressee's left), addressees were still able to use speakers' gaze cues for early disambiguation, albeit with a 150-250 ms delay due to having to re-map the speaker's non-congruent eye gaze.

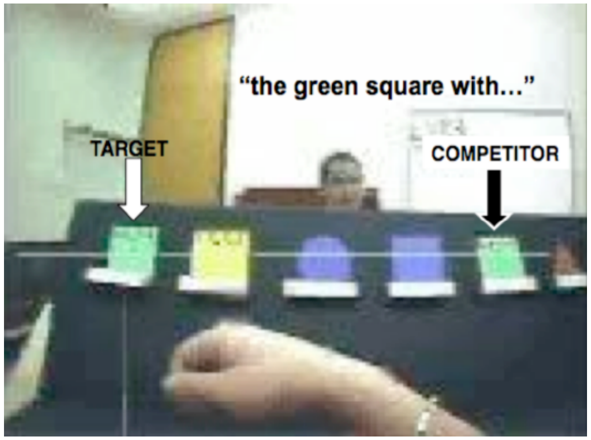

figure 1. View from the matcher's point of view (her early fixation on the target is shown by a large white crosshair). The director, who sees a display that mirrors the matcher's, is visible over the divider and is gazing at the target on her side. At this point, the speaker's referring expression is still linguistically ambiguous, but the matcher has already located it and is reaching for it.

A follow up study (Hanna, Brennan, & Savietta, 2011) used this paradigm to tease out the effects of head orientation from eye gaze by comparing a condition with speakers wearing mirrored sunglasses to one with the speakers' eyes visible. We found that the head orientation cues available in the sunglasses condition were less informative (and were monitored far less by addressees) than the speakers' eye gaze cues, and in fact head orientation cues incurred a cost when competitor objects were located close together, whereas when competitor objects were located at far ends of the display, there was an early advantage for eye gaze cues followed by a later cost (proportions of looks to the target began to increase well before the linguistic point of disambiguation, but rose more slowly later on). This cost, which we attribute to addressees following speakers' looks back and forth between same-color objects, was not present in the sunglasses condition, where proportions of looks to the target *began* to rise later, but later rose much more steeply at the point of linguistic disambiguation. We conclude from this that monitoring a partner's eye gaze can have both benefits and costs (Hanna, Brennan, & Savietta, 2011).

In this paradigm, one candidate for a gaze signature consists of repeated looks back and forth between two objects. This pattern should be interpreted differently depending on participants' roles and purpose; if the looks are by an addressee, they may mark the weighing of two potential referents for an ambiguous referring expression, whereas if the looks are by the speaker, they may signal a decision unfolding about which of two potentially relevant objects to refer to, or about how to distinguish figure from ground in planning a referring expression.

One interesting psychological issue concerning gaze signatures is whether or not a diagnostic pattern of looks is simply instrumental, that is, required for doing the task itself, as opposed to intended to be recognized as a communicative signal, such as pointing. To be communicative, a signal must have 3 characteristics (Brennan & Williams, 1995): **(1) it must be**

informative, (2) the information must be able to be perceived and processed by an addressee, and **(3) the signal must be able to be adapted by a speaker based on the speaker's intentions.** The distinction between instrumental/informative and communicative cues corresponds to Grice's distinction between natural and non-natural meaning (Grice, 1957; 1975), in which true communication involves recognition of a partner's intention to communicate. It is not always easy to tell whether a cue is used communicatively: newborn infants cry when in distress and parents recognize this (satisfying criteria 1 and 2); however the crying is not truly communicative until the child expects or intends for the parent to recognize the distress. Likewise, smoke "means" fire in the sense that it is natural consequence of fire; a smoke signal, on the other hand, is communicative when used to send a message. In collaborative tasks with a spatial component, a gaze signature or pattern of looking can be informative about what the gazer is doing or thinking without being intended to be communicative; that is probably the case for this face-to-face paradigm. On the other hand, a gaze signature may at times be expected to be recognized as purposeful.

Shared Gaze: Our early DUET system

In 2005, we implemented a shared gaze system in which two remotely located partners each wore a head-mounted Eyelink II (SR Research) eye tracker and collaborated to do several time-critical search and consensus tasks (Brennan et al., 2005, 2008; Neider et al., 2010; Zelinsky et al., 2005). This system is illustrated by the schematic in Fig. 2. Each partner's fixations were displayed in real time as a moving gaze cursor superimposed over the other's screen, so each could monitor the other's gaze.

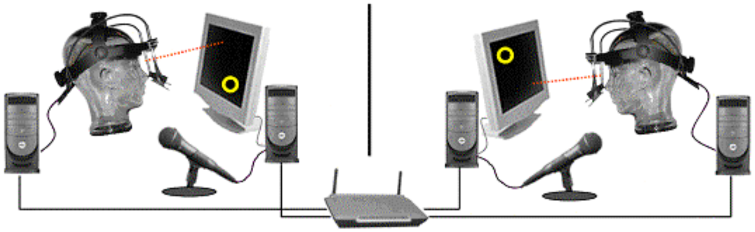

figure 2. The Shared Gaze System (Brennan et al., 2005, 2008). Each person's gaze cursor (yellow circle) is superimposed in real time on the partner's screen. There is an optional voice channel.

We compared performance and strategies for pairs searching for a target; a trial ended successfully when one member of a pair either found the target (an O among Qs) and pressed the *target present* key, or else accurately pressed the *target absent* key. There were four conditions: shared gaze with a 2-way voice channel (SG+V), shared gaze alone (SG), shared voice alone (SV), and a no communication (NC) baseline. For a simple collaborative search task (reported in Brennan et al., 2005; 2008), findings included:

- Shared gaze (SG and SG+V) cues aided joint visual search. The benefits of sharing gaze outweighed the costs of monitoring gaze cursors.
- Remarkably, collaborators could communicate strategies and coordinate behavior *using shared gaze alone (SG)*. This was the best condition overall for collaborative search. Because shared gaze cursors were displayed in the partner's own frame of reference, they proved easy and natural to use.
- Shared gaze cues (SG or SG+V) enabled spatial division of labor (Fig. 3). Whereas SV searchers divided the display coarsely (sometimes by saying "you look

left, I'll look right" and sometimes more implicitly), SG searchers used a *look where I'm not looking strategy*, which allowed for a more dynamic division of labor. Gaze also afforded a more *precise* temporal division of labor than speech (SV). Partners offered targeted assistance; if A finished her side before B, she knew from B's cursor exactly how to assist him, and he knew she was doing so.

- Adding speech (SG+V) to gaze (SG) actually *slowed* performance. This appeared to be due in part to face-management (politeness) costs, which were absent when people were not able to speak to one another (SG).

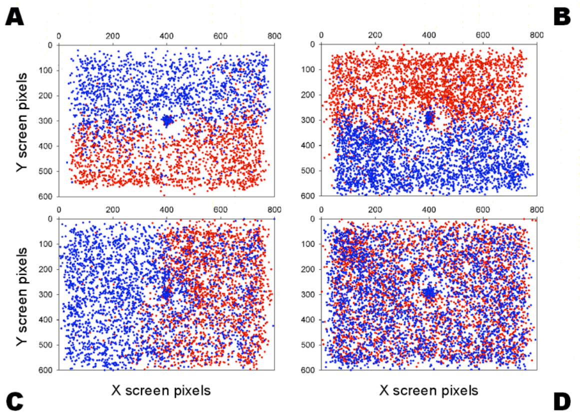

figure 3. Panels A-D: Representative fixation distributions from one pair of partners each in the SG (Panel A), SG+V (Panel B), SV (Panel C) and NC (Panel D) conditions. Fixations from one searcher are shown in red, with fixations from the other in blue. Partners divided the labor spatially in all conditions except for No Communication; what is most striking is that they did so based entirely on eye gaze cues in the SG condition (Brennan et al., 2008).

In a more complex task that required both partners to both fixate the target (the location of a sniper in an urban landscape of skyscrapers, Fig. 4) and reach consensus about it before pressing the button to end the trial, we expected that being able to use speech (SG+V) would provide a useful alerting function and aid in reaching consensus. The times for both partners to find and agree on targets were faster with shared gaze than with speech, with this benefit due primarily to faster consensus (less time needed for the second partner to fixate on the target after it was located by the first partner). In this experiment, SG+V was numerically (but not reliably) faster than SG alone. Together, our results demonstrate that sharing gaze is more efficient than speaking for the rapid communication of spatial information.

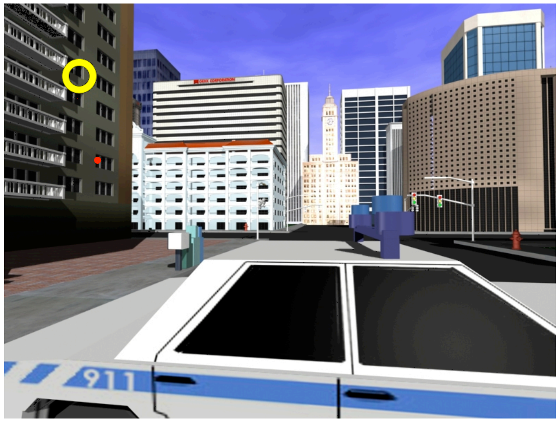

figure 4. A moving yellow ring represents the partner's gaze; the red dot below it represents the target, along with sound (a sniper firing from the window of a building).

Conclusions

Our current research includes trying to detect gaze signatures for more complex, dynamic tasks as well as for "mind reading," or inferring a goal from a searcher's eye movements (Brennan, Zelinsky). Our experiments in this program involve pairs of people as well as systems that interact with people to use their eye gaze in order to augment human performance in visual tasks. We will also use avatars as collaborative partners in order to experimentally control a partner's gaze cues (Hanna; see also Staudte & Crocker, 2011). We predict that as people begin to use DUET systems for collaborative tasks outside of the laboratory, they may learn to monitor and recognize their partners' patterns of gaze as informative (whether explicitly or implicitly).

However, an interesting question that remains is the extent to which conventions for gaze signatures may evolve for intentional (communicative) signaling among DUET users. Cues such as eye gaze and head orientation sometimes rise to the level of explicit awareness and may be consciously controlled with varying degrees of success, as some of us have discovered by observing students in lecture halls who sit facing forward while surreptitiously glancing sideways at a neighbor's exam, or as we ourselves have learned as we try to glance undetected at our watch during office hours. Anecdotally at least, it seems that people are aware that their head orientation is noticeable to others, even if they sometimes act as if their eye gaze is not.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. IIS-0713287, IIS-1111047, IIS-0527585, and ITR-0082602. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Brennan, S. E. (2004). How conversation is shaped by visual and spoken evidence. In J. Trueswell & M. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-action traditions* (pp. 95-129). Cambridge, MA: MIT Press.

[2] Brennan, S., Dickinson, C., Chen, X., Neider, M., & Zelinsky, G. (2005). When eyegaze speaks louder than words: The advantages of shared gaze for coordinating a collaborative search task. *Abstracts of the 46th Annual Meeting of the Psychonomic Society*, p. 4. Toronto, Canada.

[3] Brennan, S. E., Chen, X., Dickinson, C., Neider, M., & Zelinsky, G. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition, 106*, 1465-1477.

[4] Brennan, S. E. & Hanna, J. E. (2009). Partner-specific adaptation in dialogue. *Topics in Cognitive Science (Special Issue on Joint Action), 1*, 274-291.

[5] Brennan, S. E., & Hulteen, E. (1995). Interaction and feedback in a spoken language system: A

theoretical framework. *Knowledge-Based Systems, 8,* 143-151.

[6] Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language, 34,* 383-398.

[7] Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Washington, DC: APA. Reprinted in R. M. Baecker (Ed.), *Groupware and computer-supported cooperative work: Assisting human-human collaboration* (pp. 222-233). San Mateo, CA: Morgan Kaufman Publishers, Inc.

[8] Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge, England: Cambridge University Press.

[9] Clark, H. H., and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13,* 259-294.

[10] Grice, H. P. (1957). Meaning. *Philosophical Review, 66,* 377-388.

[11] Grice, H. P. (1989). Meaning revisited. In *Studies in the Way of Words*, 283-303. Cambridge, MA: Harvard University Press.

[12] Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language, 57*, 596-615.

[13] Hanna, J. E., Brennan, S. E., & Savietta (2011). Benefits and costs of using eye gaze in face-to-face conversation. Unpublished manuscript.

[14] McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* Chicago: University of Chicago Press.

[15] Neider, M. B., Chen, X., Dickinson, C. A., Brennan, S. E., & Zelinsky. G. J. (2010). Coordinating spatial referencing using shared gaze. *Psychonomic Bulletin & Review, 17*, 718-724.

[16] Maria Staudte and Matthew W. Crocker (2011). Investigating Joint Attention Mechanisms through Spoken Human-Robot Interaction", Cognition, 120, 268-291.

[17] Zelinsky, G., Dickinson, C., Chen, X., Neider, M., & Brennan, S. E. (2005). Collaborative search using shared eye gaze. *Abstracts of the 5th Annual Meeting of the Vision Sciences Society* (p. 115). Sarasota, Florida.