In *Lecture Notes in Computer Science* Vol. 1948, pp 1-7, Springer
- Proceedings of The third International Conference on Multimodal Interfaces, Oct 14-16, 2000, Beijing, China.

Gaze and Speech in Attentive User Interfaces

Paul P. Maglio, Teenie Matlock, Christopher S. Campbell, Shumin Zhai, and Barton A. Smith

IBM Almaden Research Center
650 Harry Rd
San Jose, CA 95120 USA
{pmaglio, tmatlock, ccampbel, zhai, basmith}@almaden.ibm.com

Abstract. The trend toward pervasive computing necessitates finding and implementing appropriate ways for users to interact with devices. We believe the future of interaction with pervasive devices lies in *attentive user interfaces*, systems that pay attention to what users do so that they can attend to what users need. Such systems track user behavior, model user interests, and anticipate user desires and actions. In addition to developing technologies that support attentive user interfaces, and applications or scenarios that use attentive user interfaces, there is the problem of evaluating the utility of the attentive approach. With this last point in mind, we observed users in an "office of the future", where information is accessed on displays via verbal commands. Based on users' verbal data and eye-gaze patterns, our results suggest people naturally address individual devices rather than the office as a whole.

1 Introduction

It is a popular belief that computer technology will soon move beyond the personal computer (PC). Computing will no longer be driven by desktop or laptop computers, but will occur across numerous "information appliances" that will be specialized for individual jobs and pervade in our everyday environment [9]. If point-and-click graphical user interfaces (GUI) have enabled wide use of PCs, what will be the paradigm for interaction with pervasive computers? One possible approach is *attentive user interfaces* (AUI), that is user interfaces to computational systems that *attend* to user actions—monitoring users through sensing mechanisms, such as computer vision and speech recognition —so that they can *attend* to user needs—anticipating users by delivering appropriate information before it is explicitly requested (see also [7]).

More precisely, attentive user interfaces (a) monitor user behavior, (b) model user goals and interests, (c) anticipate user needs, and (d) provide users with information, and (e) interact with users. User behavior might be monitored, for example, by video cameras to watch for certain sorts of user actions such eye movements [5,14] or hand gestures [1], by microphones to listen for speech or other sounds [10], or by a computer's operating system to track keystrokes, mouse input, and application use [4,6,7]. User goals and interests might be modeled using Bayesian networks [4], predefined knowledge structures [11], or heuristics [7]. User needs might be anticipated by modeling task demands [11]. Information might be delivered to users by speech or by text [7,10], and users might interact directly through eye gaze, gestures or speech [1,5,12,14].

Attentive user interfaces are related to perceptual user interfaces (PUI), which incorporate multimodal input, multimedia output, and human-like perceptual capabilities to create systems with natural human-computer interactions [10,13]. Whereas the emphasis of PUI is on coordinating *perception* in human and machine, the emphasis of AUI is on directing *attention* in human and machine. For a system to attend to a user, it must not only perceive the user but it must also anticipate the user. The key lies not in how it picks up information from the user or how it displays information to the user; rather, the key lies in how the user is modeled and what inferences are made about the user.

This paper reports the first of a series of studies investigating issues in AUI. Critical to AUI is multimodal input: speech, gesture, eye-gaze, and so on. How do people use speech when addressing pervasive computing devices and environments? Do people talk to devices differently than they talk to other people? Do people address each device as a separate social entity. Do they address the whole environment as a single entity? In the famous science fiction movie, *2001: A Space Odyssey*, the astronauts address HAL, the computer that runs the spaceship, directly; they do not address the individual objects and devices on board the ship. Is it really more natural to say "Open the pod-bay doors, HAL" than it is to say "Pod-bay doors, open"? In addition, eye gaze plays an important role in human-human communication;

In *Lecture Notes in Computer Science* Vol. 1948, pp 1-7, Springer
- Proceedings of The third International Conference on Multimodal Interfaces, Oct 14-16, 2000, Beijing, China.

for instance, a speaker normally focuses attention on the listener by looking, and looking away or avoiding eye-contact may reflect hesitation, embarrassment, or shyness. Thus, it is natural to ask, what is the role of eye gaze when speaking in an attentive environment? Will people tend to look at individual devices when communicating with them? If so, eye gaze might provide additional information to disambiguate speech.

2 Experiment

We investigated how people naturally interact with an " office of the future" . To separate conceptual issues from current technology limitations, a Wizard-of-Oz design was used to provide accurate and timely reactions to user commands and actions. The devices in the office were controlled by one of the experimenters hidden behind a wall. Users were given a script containing a set of office tasks and were told to perform them using verbal commands only. A green blinking light served as feedback that the command was understood and being executed. There was one between-subjects factor with two levels, distributed feedback and non-distributed feedback. In the distributed condition (DC), feedback in the form of the green flashing light was seen on each device. In the non-distributed condition (NC), feedback appeared in a single location on the wall representing the " office" . Non-distributed feedback was meant to suggest that the office itself process the commands (kind of like talking to HAL in 2001), whereas distributed feedback was meant to suggest that individual devices process the commands (unlike talking to HAL). We were interested in whether and how people's behavior—verbal and gaze—might change with the kind of feedback provided.

2.1 Method

Thirteen volunteers were randomly placed into one of two conditions—DC or NC. In both, participants were given a script in the form of a list containing six tasks, such as get an address, dictate memo, print memo, find date from calendar, get directions, and print directions. These were to be completed using four devices available in the room: an address book, a calendar, a map, and a dictation device. Neutral language in the instructions (e.g., " You will need to make a printout of the document") was used so as not to bias participants' utterances toward giving command to individual devices (distributed) or to the room as a whole (non-distributed). As a cover story, participants were told that a wide range of voices were needed to test IBM's "office of the future" speech recognition project. This way, participants were not aware we would be looking at the nature of their verbal commands, and would feel less self-conscious about the types of language they were generating. It also ensured that participants would speak loudly and clearly. Participants were told to take their time and to repeat verbal commands if they encountered a lag or a problem in being understood. Hidden cameras on the devices recorded gaze information, and a microphone recorded verbal commands.

2.2 Attentive Office

The attentive office contained three small flat screen displays were labeled "Calendar" , " Map/Directions", and " Address" , and a plastic, futuristic-looking orb was labeled "Dictation" . There was also a printer, and a large flat screen display without a function, but with futuristic images displayed on it (simply meant to distract user's attention). All displays were 800x600 pixel LCD flat panels. In the DC, a small black box with a green light (feedback module) was attached to the top of each screen. For the dictation device, no screen was used, so the feedback module was placed behind the futuristic orb. No traditional manual input devices (keyboards, mice, or other control buttons) were in the room. During the experiment, devices displayed the requested information immediately after the request. The information appeared in a futuristic-looking graphical display that did not look like the typical Windows desktop environment (see Figure 1).

In *Lecture Notes in Computer Science* Vol. 1948, pp 1-7, Springer
- Proceedings of The third International Conference on Multimodal Interfaces, Oct 14-16, 2000, Beijing, China.

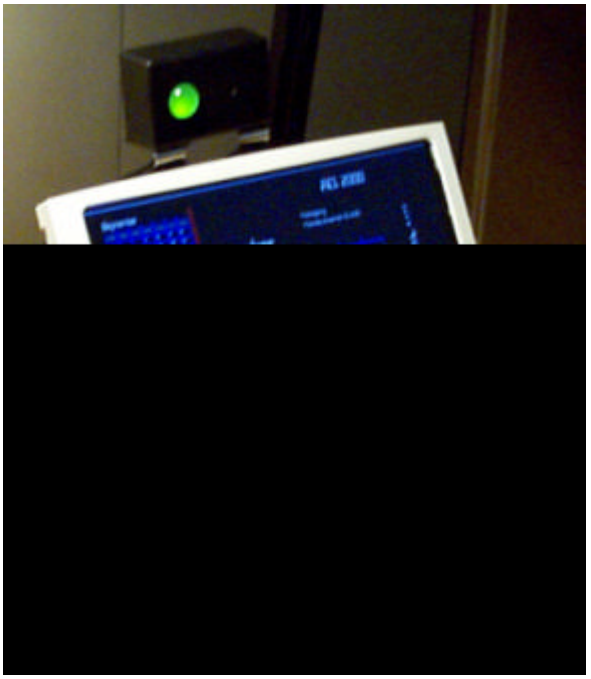

Fig 1. Futuristic display used in the attentive office.

3 RESULTS

The language data were coded as follows. Verbal commands were first placed into four categories based on type of request: imperative, subject noun-phrase, question, and other. An imperative request was a direct command to perform an action (e.g., "Get directions to Dr. Wenger's."). A subject noun-phrase request was a first-person statement highlighting the requestor's goal or desire (e.g., "I want directions to Dr Wenger's home"). A question was a was an interrogative statement requesting an action (e.g., "How do I get to Dr. Wenger's home?"). Finally, remaining requests were put in the other category, including fragmented requests (e.g., "Wenger").

For each type of command, utterances were then divided into one of two categories depending on whether the participant specified an addressee when giving the command. Specified addressee requests included a reference to agent or actor (e.g., " Printer, print memo" and " Calendar, when is Thanksgiving 2000?"), and non-specified addressee requests did not explicitly mention the addressee (e.g., " Print memo" and " When is Thanksgiving 2000?").

For the verbal commands, results indicate that participants made many more imperative requests (62%) than question requests (17%), subject noun-phrase requests (13%), or other requests (8%). Figure 2 shows the proportion of requests in each category for DC and NC (differences between DC and NC were not reliable within each of the four categories). The language data also show that for all verbal commands (collapsed across category), very few of the requests were specified (< 2%). That is, only a few instances of statements such as, "Printer, give me a copy" , emerged in the data. Conversely, nearly all commands featured no specific addressee, such as " Give me a copy" .

Gaze patterns were coded according to if and when participants looked at the devices or the "room" (the light on the wall in NC). Namely, for each utterance, the participant either looked at the appropriate device or at the wall before (or during) speaking or after speaking. Figure 3 shows proportion of gazes occurring before/during, after, or never in relation to the verbal request. Overall, the data show more looks occurred before requests than after. In addition, participants

nearly always looked at the device when making a request. As with types of requests made, gaze times did not differ for DC and NC.

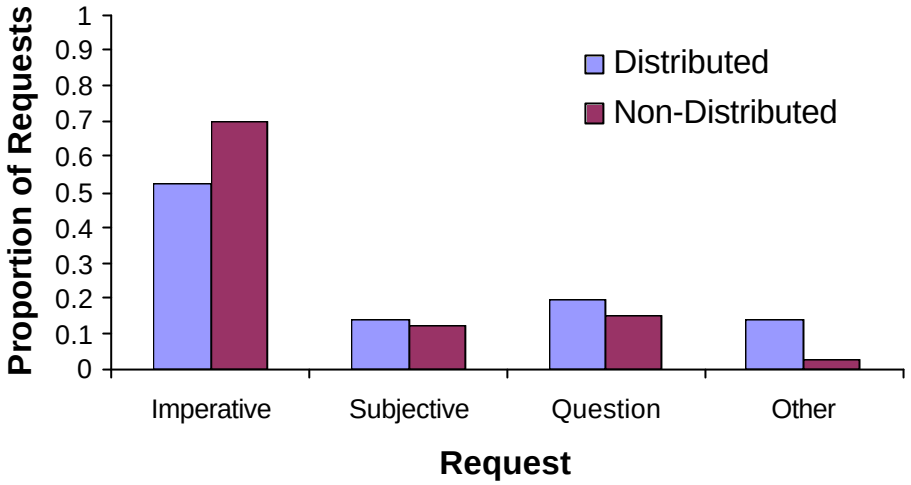

Fig 2. Proportion of utterances in the four categories.

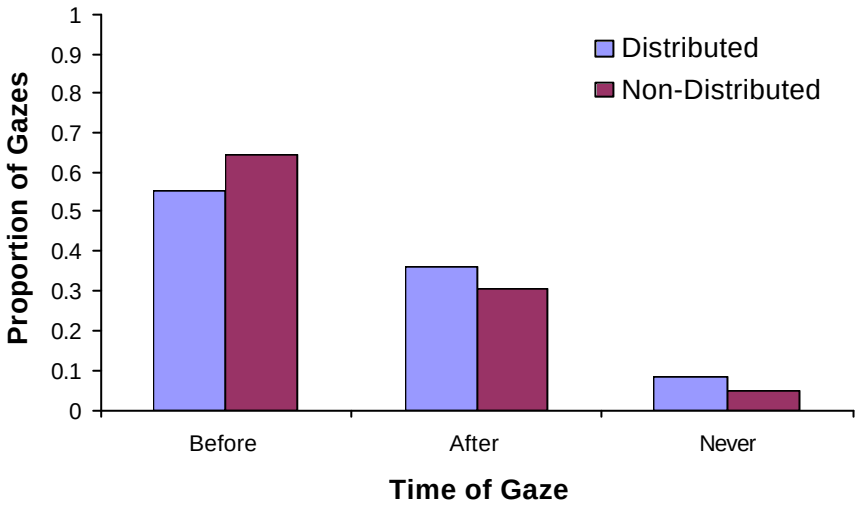

Fig 3. The proportions of gazes at the device before and after speaking.

4 Discussion

We explored the nature of user interaction in an attentive office setting. The pattern of results we found suggests that the DC environment was more natural for participants than was the NC environment Though most verbal requests did not address specific devices, the majority of looks at devices occurred before the requests. Looking before speaking is seen in human-human communication: People often use eye gaze to establish or maintain social contact or interaction.

Thus, looking before speaking may indicate that participants were specifying the recipient of the request through eye gaze, as if they default to a natural social interaction with individual devices.

Interestingly, the pattern of results—both verbal and gaze data—did not vary with experimental condition. There was no fundamental difference between the way people spoke to or directed gaze in the DC or NC environments. Our expectation prior to performing the study, however, was that in NC, centralized feedback (via a single flashing light) might bias participants not to look at individual devices, or might encourage participants to verbally specify an addressee more often. There are at least three possible explanations for why no differences arose in language and gaze patterns between the two conditions. First, multiple devices were present in both conditions. The mere presence of multiple devices in the non-distributed condition may have influenced people to interact with each device individually rather than with the single flashing light on the wall. Second, the flashing lights may not have provided a compelling a source of feedback, especially in the non-distributed condition. Third, the role of the flashing light may have been unclear in general. For instance, did it refer to an assistant, an authority, a monitor, and so on? In future studies, we plan to address this last issue by varying the script across conditions to highlight the different possible roles of the agent providing feedback.

In many voice-activated user interfaces, verbal requests typically must be spoken in a scripted and unnatural manner so that the recipient of the message can be easily determined [1]. Our data suggest that gaze information alone will disambiguate the recipient 98% of the time. For example, if one says " Wenger's home" while looking at the address book, it is clear that the address for Wenger's home is desired. By automating the attentive office to coordinate voice and gaze information in this way, a user's tacit knowledge about interpersonal communication can enable natural and efficient interactions with an attentive user interface. The preliminary study reported here marks only the first step toward understanding the ways in which people naturally interact with attentive devices, objects, and environments.

Acknowledgments

Thanks to Dave Koons, Cameron Miner, Myron Flickner, Chris Dryer, Jim Spohrer and Ted Selker for sharing ideas and providing support on this project.

References

1 Bolt, R. A. (1980). Put that there: Voice and gesture at the graphics interface. *ACM Computer Graphics 14(3)*, 262-270.

2 Coen, M. H. (1998). Design principles for intelligent environments, In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI '98)*. Madison, WI.

3 Hirsh, H., Coen, M.H., Mozer, M.C., Hasha, R. & others. (1999). Room service, AI-style. *IEEE Intelligent Systems, 14(2),* 8-19.

4 Horvitz, E. Breese, J., Heckerman, D., Hovel, D., &. Rommelse, K. (1998). The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users, in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 256-265.

5 Jacob, R. J. K. (1993). Eye movement-based human-computer interaction techniques: Toward non-command interfaces. In Hartson, D. & Hix, (Eds.)., *Advances in Human-Computer Interaction, Vol 4*, pp. 151-180. Ablex: Norwood, NJ.

6 Linton, F., Joy, D., & Schaefer, H. (1999). Building user and expert models by long-term observation of application usage, in *Proceedings of the Seventh International Conference on User Modeling*, 129-138.

7 Maglio, P. P, Barrett, R., Campbell, C. S., Selker, T. (2000) Suitor: An attentive information system, in *Proceedings of the International Conference on Intelligent User Interfaces 2000*.

8 Maglio, P. P. & Campbell, C. S. (2000). Tradeoffs in displaying peripheral information, in *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2000)*.

9 Norman, D. A. (1998). *The invisible computer*. Cambridge, MA: MIT Press.

10 Oviatt, S. & Cohen, P. (2000). Multimodal interfaces that process what comes naturally. *Communications of the ACM, 43(3)*, 45-53.

11 Selker, T. (1994). COACH: A teaching agent that learns, *Communications of the ACM, 37(1),* 92-99.

12 Starker, I. & Bolt, R. A. A gaze-responsive self-disclosing display, in *Proceedings of the Conference on Human Factors in Computing Systems, CHI '90*, 1990, 3-9.

13 Turk, M. & Robertson, G. (2000). Perceptual user interfaces. *Communications of the ACM, 43(3)*, 33-34.

14 Zhai, S., Morimoto, C., & Ihde, S. (1999). Manual and gaze input cascaded (MAGIC) pointing, in *Proceedings of the Conference on Human Factors in Computing Systems (CHI 1999)*, 246-253.