

See What I Mean: On Gaze and Speech in Attentive Environments

Teenie Matlock Christopher S. Campbell Paul P. Maglio Shumin Zhai Barton A. Smith
IBM Almaden Research Center
650 Harry Rd
San Jose, CA 95120 USA
{tmatlock, ccampbel, pmaglio, zhai, basmith}@almaden.ibm.com

ABSTRACT

The trend toward pervasive computing necessitates finding appropriate ways for users to interact with devices. With this in mind, we observed users in an “office of the future”, where information is accessed on displays via verbal commands. Results suggest people naturally attend to individual devices rather than to the office as a whole.

Keywords

Attentive User Interface (AUI), intelligent environments, multimodal input, information appliances.

INTRODUCTION

It has become a popular belief that computer technology will soon move beyond the personal computer (PC). Computing will no longer be concentrated around desktop or laptop PCs, but will occur over numerous specialized “information appliances” that will pervade our work and everyday environments [4]. The point-and-click Graphical User Interface (GUI) has enabled the general population to use PCs, but in a future with no central screen and keyboard, what will be the paradigm for interaction with pervasive computers? Though there is some user interface work on pervasive computing and intelligent environments (e.g., [2]), a number of critical issues have yet to be addressed.

We envision three broad categories of interaction methods in the post-PC era: (a) GUIs, in appliances or in a control console, which might be hand-held and wireless; (b) physical control buttons or sliders on appliances, not unlike today’s VCRs; and (c) Attentive User Interfaces (AUI), systems or environments that monitor users through sensing mechanisms, such as computer vision and speech recognition (see also [3]). The latter interaction type has long been featured in sci-fi, from “2001” to “Star Trek”. We report results from a preliminary study that explores basic issues of the AUI paradigm.

Our study was conducted in an “office of the future” mock up. The goal here is not the specific design of the mock up, but rather a basic question about AUI environments: Do people *expect* to interact with many devices individually or do they expect to interact with the office as a single entity? We are also interested in the types of speech people use and

where people look when speaking in the AUI environment.

EXPERIMENT

To separate conceptual issues from current technology limitations, a Wizard-of-Oz design was used to provide accurate and timely reactions to user commands. The reactions of the office were controlled by one of the experimenters behind a wall. Users were given a set of office-related tasks to perform using verbal commands. 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 (DF) and non-distributed feedback (NF). In the DF condition, feedback in the form of a green flashing light was seen on each device. In the NF condition, feedback appeared in a single location on the wall, representing the “room”. We were interested in whether and how people’s behavior—verbal and gaze—change with the kind of feedback provided.

Method

Thirteen volunteers were randomly placed into one of two conditions—six in DF and seven in NF. In both, participants were given a list of seven tasks, such as get address, dictate memo, print memo, find date from calendar, get directions, and print directions. These were to be completed using devices such as an address book, a calendar, a map, and a dictation device. Hidden cameras on the devices recorded gaze information, and a microphone recorded verbal commands. As a cover story, participants were told that a wide range of voices were needed for IBM’s “office of the future” speech recognition project. In the instructions, neutral language was used so as not to bias participants’ utterances toward distributed or non-distributed commands.

Attentive Office

The attentive office was set up as follows. Three screens 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 with futuristic images displayed on it (meant to distract users’ attention). All displays were 800x600 pixel LCD flat panels. In the DF condition, 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 orb. No traditional manual input devices (keyboards, mice, or other control buttons) were in the room. In the NF condition, a single feedback module was affixed to the wall rather than to the devices.

During the experiment, devices displayed the requested information immediately after the command. Information appeared in a futuristic-looking graphical display that did not look like the typical Windows desktop environment.

RESULTS

Utterances were put into four categories based on type of request: imperative, subject noun-phrase, question, and other. An imperative was a direct command to perform an action (e.g., "Get directions to Dr. Wenger's."). A subject noun-phrase was a statement of requestor goal or desire (e.g., "I want directions to Dr Wenger's home"). A question was a request for information (e.g., "Where is Dr. Wenger's home?"). Finally, the other category including fragmented requests (e.g., "Wenger's address"). Utterances were also divided into specified-addressee and non-specified-addressee. Specified requests included an explicit reference to agent or actor (e.g., "Printer, print memo.") and non-specified requests did not (e.g., "Print memo").

Participants made many more imperative requests (62%) than question requests (17%), subject noun-phrase requests (13%), or other requests (8%). Another striking finding was that very few of the requests were specified (<2%). Figure 1 shows proportion of requests in each category for DF and NF. Apparently, DF versus NF visual feedback did not affect verbal commands.

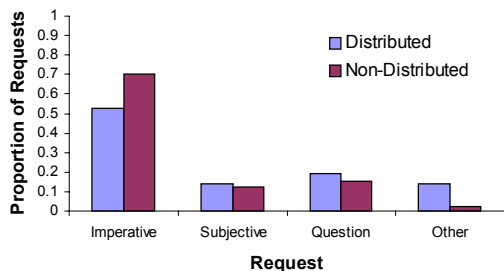


Figure 1: Proportion of utterances in the four categories.

Gaze patterns were coded according to if and when participants looked at devices or "room" (the light on the wall in NF). For each utterance, the participant either looked at the appropriate device or at the wall before (actually, before or during) speaking or after speaking. Figure 2 shows proportion of gazes that occurred before, after, or never in relation to verbal request. Overall, more looks occurred before requests than after. Participants also nearly always looked at the device when making a request. As with types of requests, gaze times did not differ for DF and NF.

DISCUSSION

In this preliminary study, we explored the nature of user interaction with attentive environments. Though most verbal requests were non-specified—did not explicitly address specific devices or the room—the majority of looks occurred before the verbal requests. Looking before speaking may indicate participants were using gaze to address the recipient of the request, as looking before speaking is also seen in human face-to-face communication.

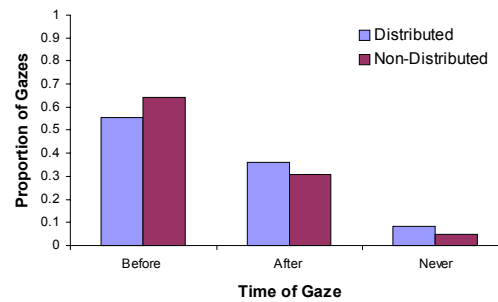


Figure 2: The proportions of gazes at the device before and after speaking.

This observed pattern in the attentive office suggests participants default to interacting with individual devices. Thus, our results suggest that the DF environment was more natural for participants than was the NF environment.

Interestingly, verbal and gaze patterns did not vary with experimental condition. In NF, centralized feedback could have biased the participants not to look at individual devices and to use verbal addressing more frequently, but it did not. There are at least two possible conclusions: (a) the effect is too strong to be manipulated by feedback placement, or (b) type of feedback (blinking light) was not compelling.

In voice-activated attentive environments, verbal requests typically must be spoken in a scripted manner so the recipient of the message can be determined [1]. But our data suggest that gaze information 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 post-PC computing environments to account for voice and gaze information, existing knowledge about interpersonal communication can be used to create natural and efficient attentive user interfaces.

ACKNOWLEDGMENTS

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

REFERENCES

1. Coen, M. H. (1998). Design principles for intelligent environments, In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*. (AAAI'98). Madison, WI.
2. Hirsh, H., Coen, M.H., Mozer, M.C., Hasha, R. & others. (1999). Room service, AI-style. *IEEE Intelligent Systems*, 14(2), 8-19.
3. Maglio, 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*.
4. Norman, D. A. (1998). *The invisible computer*. Cambridge, MA: MIT Press.