On Understanding Discourse in Human-Computer Interaction

Paul P. Maglio (pmaglio@almaden.ibm.com)
Teenie Matlock (tmatlock@psych.stanford.edu)
Sydney J. Gould (sydneygould@hotmail.com)
Dave Koons (dkoons@almaden.ibm.com)
Christopher S. Campbell (ccampbel@almaden.ibm.com)
IBM Almaden Research Center
650 Harry Rd, B2-NWE
San Jose, CA 95120 USA

Abstract

We report on an experiment that investigated how people naturally communicate with computational devices using speech and gaze. Our approach follows from the idea that human-human conversation involves the establishment of common ground, the use of gaze direction to indicate attention and turn-taking, and awareness of other's knowledge and abilities. Our goal is to determine whether it is easier to communicate with several devices, each with its own specialized functions and abilities, or with a single system that can control several devices. If conversations with devices resemble conversations with people, we would expect interaction with several devices to require extra effort—both in building common ground and in specifying turn-taking. To test this, we observed participants in an office mock-up where information was accessed on displays through speech input only. Between groups, we manipulated what participants were told: in one case, that they were speaking to a single controlling system, and in the other, that they were speaking to a set of individually controlled devices. Based on language use and gaze patterns, our results suggest that the office environment was more efficient and easier to use when participants believed they were talking to a single system than when they believed they were talking to a several devices.

Introduction

One approach to human computer interaction is to improve the usability, user experience, and intuitiveness of technology by creating natural user interfaces. Here, *natural* refers to interactions that are like those people have with one another. Such is the goal of multimodal or attentive systems (Maglio, Matlock, Campbell, Zhai & Smith, 2000; Oviatt & Cohen, 2000), and speech and conversational interfaces (Maybury, 1997). Understanding cues in conversation, language use, perceptual abilities, and expectations is vital to building systems that can be used with little training.

Advances in technology are resulting in smaller, cheaper, and more pervasive computational systems than ever before. But are we ready for this surge of electronics and information? No longer confined to desktop or laptop machines, computational systems will soon extend across numerous "information appliances" that are specialized for individual jobs and embedded in the everyday environment (Norman, 1998). If point-and-click graphical user interfaces (GUI) have enabled wide use of PCs, what will be the paradigm for interaction with pervasive computing systems? As natural human-computer interfaces and pervasive systems converge, what form will technology take?

To address these questions, we explored the design of a pervasive system with speech input in an office setting. We were concerned specifically with conversational cues that people rely on when interacting with the system. Some evidence suggests that people can attribute human-like or social qualities to computers with which they interact; for instance, networked computers described as physically close to the user are judged as more helpful than those described as physically distant (Reeves & Nass, 1996). Although people do not treat computers as true conversational partners (Yankelovich, Levow & Marx, 1995), these sorts of results suggest that people apply natural ways of interacting to situations in which the conversational partner is a computer or other computational device.

Our main concern is whether it is easier for people to talk to a single system or to a collection of devices. In a previous study of a speech-controlled office, we found behaviors and attitudes depended on whether users received simple command recognition feedback (a blinking light) from the various devices that performed tasks or from a single, central location (Maglio, Matlock, Campbell, Zhai & Smith, 2000; Matlock, Campbell, Maglio, Zhai & Smith, 2001). In that study, users were faced with simple office tasks (such as looking up information, dictating a letter, and printing a letter) to be completed using speech input only. To do this, users were given a set of physical displays dedicated to various functions (such as address book, calendar, and so on). Between groups of participants, we manipulated whether feedback was associated with individual displays or with the room as whole. This feedback manipulation was meant to suggest either central control or distributed control. Behaviorally, we found that regardless of condition, participants rarely addressed individual devices verbally, but they looked at the devices that they expected to display the results

before they spoke (Maglio et al., 2000). In a questionnaire aimed at uncovering attitudes toward the office, we found that participants in the central condition were more likely to rate their interactions with the office as being similar to interactions with people than were those in the distributed condition (Matlock et al., 2001). The results show that although people judge the central controller to be more like a person, they interact with devices individually in both cases, looking at devices when they speak. One design implication is that the feedback provided by blinking lights enables natural user-computer interactions. But the question of whether it is easier to speak to a single system or to multiple devices remains.

Let us first consider how people use language to communicate. There are many theories. A popular view is that discourse is a shared activity whereby two or more individuals cooperate to build and achieve understanding (Clark, 1996). This joint activity view implies that the meaning of an utterance is determined not only by what the speaker wishes to say to the listener, but also by context. This includes speaker's beliefs about the situation, (e.g., what speaker assumes the listener knows about the context), common ground (e.g., shared history), and listener's ability to accurately interpret the speaker's message (e.g., listener is paying attention). For example, imagine that it's early afternoon and you have just come back from a favorite lunch spot. A friend looks at you and asks, "Was it crowded?", where *it* refers to the restaurant. It is no problem to use the indexical *it* because the friend can assume that you know which restaurant is being asked about. The question can even be reduced further by simply asking, "Crowded?", and you are still likely to understand what is meant. This type of coordinated interaction is so common and natural that people do not think twice about it.

Given context's role in understanding, the joint activity view implies that the process of conversation also involves verbal (e.g., prosodic) and non-verbal (e.g., gaze) cues to convey meaning. Feedback is critical for supporting user interactions with computational systems (Perez-Quinones & Sibert, 1996); for instance, appropriate acknowledgments (e.g., "uh-huh") based on prosodic cues in users' speech can improve user evaluation of the system (Tsukahara & Ward, 2001). Likewise, gaze provides important cues to attention and turn-taking in group interactions (Kendon, 1967; Argyle & Cook, 1976).

Hypotheses

Following the joint activity view of conversation, our main hypothesis is that given complex tasks and their dependencies on one another, participants who interact with a single system will be more likely to establish context and then assume the system shares it than will those who interact with several devices. In human-human communication, the number of words used by participants in a conversation decreases over time, suggesting that the more common ground, the easier the communication (Clark & Wilkes-Gibbs, 1986). Thus, if people treat the system in the central condition in some ways like a single other person and they treat each of the devices in the distributed condition in some ways like several other people, participants may feel that they need to establish common ground with the single system only once but that they need to establish common ground with each of the devices individually, and we can measure this by counting words used—specifically, words that refer to things mentioned previously. Moreover, if establishing common ground is easier in the central condition than it is in the distributed condition, many predictions follow; for instance, participants should make fewer mistakes in the central, and participants should be more engaged with the central system, as shown by gaze.

An alternative to the joint activity view is that discourse is simply a process of transferring information without reference to context or to those involved. On this view, meaning is derived from what is spoken without concern for who the speaker is or what the situation is. If this is the case, maintaining separate functions in separate devices might make it easier for users to keep the various functions straight, as each device naturally conveys its own range of available options (e.g., email device for email). Thus, on this view, common ground is not constructed over time but is established once by what the device can do. If this is the case, talking to a single system ought to be more difficult than talking to multiple devices, as the single system does not make the options apparent.

Experiment

The goal of the experiment was to investigate whether and how language use and gaze would differ between participants interacting with a central system and those interacting with a distributed system. Our study was done in a mock office in which participants completed office tasks under the illusion that they were controlling what was displayed on four specialized screens. This sort of mock-up or *Wizard-of-Oz* method is often used to investigate user expectations and performance with speech-based systems (Dählback, Jönsson & Ahrenberg, 1993; Gould, Conti & Hovanyecz, 1983). The Wizard-of-Oz method relies on human controllers behind the scenes to create the appearance of an intelligent system, mocking up displays and interaction results to collect performance data.

In one condition, the experimenter instructed participants to speak to a system that controlled all the devices, and in the other, to speak to the individual devices. Unknown to participants, two other

experimenters in a separate room watched and listened, controlling what was displayed from a palette of many possible screens. The rules of the game were for the experimenters to simply behave intelligently: if what the participant was trying to do was clear from speech and other context, the system was to respond appropriately. The experimenters controlling the system were blind to which participants were given which instructions.

Method

Between two groups of participants, we manipulated *only* the instructions. In the central condition, participants were repeatedly told that they were to talk to a single computer system that displayed information on four displays. In the distributed condition, participants were repeatedly told that they were to talk to four separate information devices. Tasks were identical in both cases: sending and receiving email, updating address information, scheduling appointments, arranging a flight, and registering for a conference. Information was displayed the same way in both cases.

Participants Eighteen participants (13 females and 5 males) were recruited from summer student interns and office staff at our research lab, and paid for their time.

Materials and Apparatus Our office mock-up contained four 15-inch liquid crystal displays (LCDs) arranged on an L-shaped desk. Embedded in the bezel of two of the LCDs were pinhole video cameras, which enabled eye gaze and body position to be easily recorded. A third camera mounted on the wall above the room recorded an overview of the scene. Each display was dedicated to a different function or task: email, calendar, travel planning, and address book.

There were two sets of instructions, one for each condition. Instructions for the central condition told participants to talk to the "BlueSpeak system", a single computer system that controlled four displays. Instructions for the distributed condition told participants to talk to a set of "BlueSpeak devices", four separate devices that ran autonomously. In both cases, the script was written as a memo from a fictitious manager named Bob Wilson. The memo told the participant that he or she was to be his temporary assistant (or temp) for the day. It asked the temp to register Bob for a conference, add a new address to his address book, get a flight from San Jose to New York, reschedule a meeting, and request a vegetarian meal on the flight. The script purposely did not specify how to issue commands. It included statements such as "You will need to arrange my travel to New York and from San Jose", "I need to register for the XYZ Conference," and "Make sure my calendar is updated". Such language made for many possible ways of making requests. In addition, a few tasks were given in email messages sent to the temp during the course of the session. Table 1 shows the set of tasks and subtasks each participant was expected to carry out.

Table 1. Tasks completed by participants.

Update Address Book
- open new address form
- dictate name, address, city, state, zip, phone, email

Register for XYZ Conference
- find XYZ information screen
- obtain Bob's personal information
- compose new email to XYZ
  - dictate name, email, phone, credit card
- add XYZ to calendar

Find and Reserve Flights
- find airline reservation screen
- interact with reservation "system"
  - dictate cities, dates, non-stop, under $400
  - reserve/book itinerary
    - obtain Bob's personal information
    - dictate name, email, credit card
  - add flights to calendar

Reschedule Meetings
- obtain Kathy's personal information
- compose new email message to Kathy
- read Kathy's response
- modify Kathy's meeting

Notify Bob of Status
- compose new email to Bob
- read Bob's response
- adjust calendar , cancel meeting
- modify airline reservation
  - find reservation for Bob
  - specify vegetarian meal

Procedure Participants read the instruction sheet and were told their input was valuable because it would help ensure that the "BlueSpeak system" or the "BlueSpeak devices" would be tested on a wide range of voices. Participants were then taken into the "Office of the Future", and asked to test out their voices by reading a short passage to the system in the central condition or to the devices in the distributed condition. Participants were then told to carefully read the memo left by Bob Wilson. In all cases, participants were instructed to speak naturally and to do the best they could. They were told that there was no right or wrong way to speak to the system or the devices, and that if they were not understood, to try speaking differently. After issuing a command, the system did not give any feedback other than displaying the result of the request.

Results

All participants successfully completed the session. Few problems arose and on average it took participants 13 min 43 sec to complete all tasks. Data analysis targeted language-use and eye-gaze during the session. Only reliable differences are reported, except as noted.

Language Qualitatively, participants spoke to the system in a variety of ways. For instance, requests to send email included, "Let's send an email to Kathy Webster," "I need to send an email now – I would like to send it to k webster at ibm dot com," "Email k webster at ibm dot com," and "Write an email to Kathy Webster". Requests to get Bob a vegetarian meal on included, "Request vegetarian meal," "Vegetarian meal," "Let's make this a vegetarian meal", and "Special request, vegetarian meal for this flight".

More precisely, transcribed utterances in both conditions were examined for certain characteristics of language use. First, requests were placed into four categories: imperative, elliptical, first person, and question (cf. Maglio et al., 2000). Imperative requests are commands, such as, "update the address book", "view addresses", "register for conference". Elliptical requests contain no a verb, such as, "XYZ conference", "new entry", and "Kathy Webster". First person requests include either a singular or plural first person subject, such as "let's read this email", and "I want a vegetarian meal". Question requests include queries such as, "can I check my email?" and "are there any other flights available?". Figure 1 shows the break down of requests for both conditions. There were more imperatives (central, 70.4%; distributed, 80.3%; χ^2 = 9.75, $p < 0.01$) and more ellipticals (central: 11.8%; distributed: 16.9%; χ^2 = 4.16, $p < 0.05$) in the distributed condition, and there were more first persons (central, 13.3%; distributed, 2.8%; $\chi^2 = 26.9$, $p < 0.01$) and questions (central, 4.5%; distributed, 0.0%; χ^2 = 16.4, $p < 0.01$) in the central condition.

Second, we examined how participants verbally addressed individual devices. Specifically, we counted the number of times a device was specifically addressed by name, such as, "Address book, what is Kathy Webster's address?". The proportion of requests containing an addressee was greater in the distributed condition than in the central condition (central, 1.7%; distributed, 14%; $\chi^2 = 40.87$, $p < 0.01$).

Third, we examined the way participants recovered from errors. We were interested in how requests were reformulated after an initial attempt had failed. To take one example, the most problematic part of the script was ordering a vegetarian meal for the flight (last task in Table 1). Overall, in the central condition, the meal request was restated 9 times. In the distributed condition, it was restated 13 times. Three participants in the distributed condition were unable to complete this task at all and eventually gave up.

Fourth, we looked at the way participants relied on previously established context. For example, when interacting with the system, a participant might say "register Bob Wilson for XYZ conference" and then a short time later say, "add event to calendar", referring implicitly to the conference. In this case, the

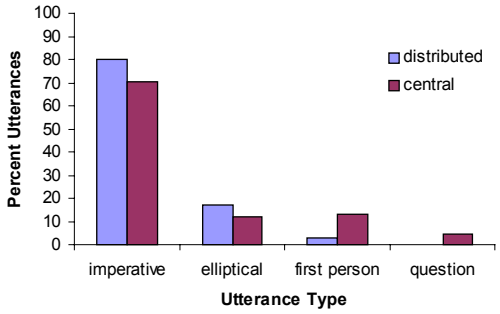

Figure 1. Percentage of the time participants used different types of requests.

participant assumes the system is following the discourse, and that once established, the context (the conference event) need not be repeated. We calculated the proportion of statements that assumed context available in a previous statement across all participants in each condition (see Figure 2). Overall, participants in the central condition assumed that the system would understand the context more often than participants in the distributed condition (central, 7.0%; distributed, 1.4%; $\chi^2 = 14.50$, $p < 0.01$).

Finally, we charted how language use changed during the course of a session. Each participant's discourse was cut in half, based on the task breakdown in Table 1. The number of times each participant relied on established context (as defined previously) was tallied separately for the first half and for the second half (see Figure 2). No difference between central and distributed conditions was found for the first half (central, 5.5%; distributed, 2.2%; $\chi^2 = 2.45$, NS), but a reliable difference was found for the second half (central, 8.5%; distributed, 0.6%; $\chi^2 = 13.59$, $p < 0.01$).

Behavior and Gaze Behaviors—including actions participants took and where they looked—were analyzed in terms of the task breakdown in Table 1. Specifically, all overt physical actions taken by participants were transcribed from the videotapes and time-stamped. From these data, we extracted number of tasks, time taken per task, number of gazes or looks to task-relevant and to task-irrelevant locations, and number and kind of the errors made. For all results, scores falling outside two standard deviations from the mean were removed and replaced by mean scores; these outliers constituted 8% of the scores.

Mean completion time was 13 min 17 sec for the central condition and 14 min 10 sec for the distributed condition. To calculate the time taken for each individual action for each participant, the time taken for each task (e.g., "Update Address Book," "Register for Conference," etc.) was divided by the number of actions (e.g., "open address book", "find Bob's personal information", etc.) actually taken to complete

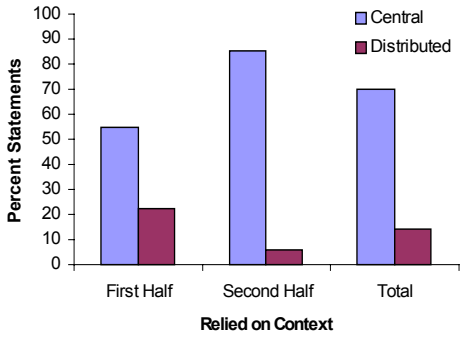

Figure 2. Percentage of time participants relied on established context.

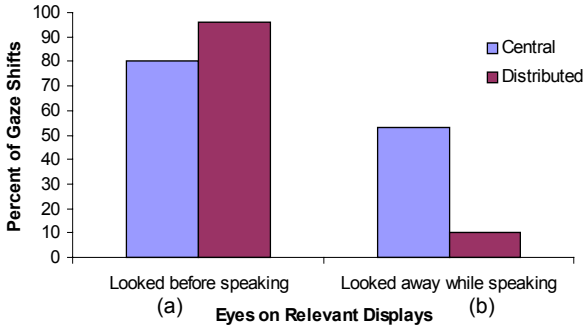

Figure 3. Percentage of time participants (a) shifted gaze to display before speaking, and (b) away from display while speaking.

the task. A difference was found between central and distributed conditions on the mean time to take an action in the address book task (central, 21.1; distributed, 27.1; $t(16) = 3.18$, $p < 0.01$), and a marginal difference was found between the mean times to take an action in the flight reservation task (central, 54.0; distributed, 73.4; $t(16) = 1.71$, $p < 0.11$).

Unnecessary actions and omitted actions constituted errors. Number of errors was calculated for each participant. Percentage of errors was determined by dividing the total number of errors by the total number of actions taken for each participant. Overall, there were 20% errors in the central condition and 16% in the distributed condition. This difference was not reliable.

Finally, we examined where people looked and when they altered their gaze. In particular, we counted the total number of times a participant looked at a specific display when making a request for which that display would be expected to show a result (see also Maglio et al., 2000). For instance, we counted times when a participant looked at the address book and then said "Michael Smith's address," but not times when the participant would say "Michael Smith's address" before looking at the address book. The percentage of the time each participant looked at the appropriate display when taking action was calculated by dividing the number of appropriate looks by the number of actions. As shown in Figure 3a, a difference was found between the two conditions (central, 80%; distributed, 96%; $t(16) = 2.79$, $p < 0.05$). In addition, we counted the number of times a participant looked away from a display they were using to complete an action (again, normalizing with respect to total number of actions). As shown in Figure 3b, a difference was found between the two conditions (central, 53%; distributed 10%; $t(16) = 2.39$, $p < 0.05$).

Discussion

In summary, participants interacting with the single system had an easier time than those interacting with multiple devices. Specifically, the data show

1. dominant use of the imperative in both cases, and more use of first-person and question forms in the central case
2. less verbal addressing in the central case
3. less reformulation—and more successful reformulation—of requests in the central case
4. more reliance on context in the central case, which increased over time
5. slightly faster overall completion time for the central case, and significantly faster for certain tasks (updating address book, reserving a flight)
6. less of the time, gaze shifted to the appropriate display before speaking in the central case
7. more of the time gaze shifted away from the appropriate display in the central case

Returning to our hypotheses, we can conclude that participants in the central condition relied more heavily on context they had established previously, shown by the number of times they implicitly referred to objects and information. This is what we expected given the joint activity view of discourse (Clark, 1996). It follows that participants in the central condition behaved more like they were speaking to a single entity than those in the distributed condition. The way participants addressed displays and shifted gaze also supports this conclusion. Participants in the central condition addressed individual devices less frequently, suggesting that they were less likely to be speaking directly to devices. Moreover, participants in the central condition shifted gaze to individual devices when starting a task less frequently, also suggesting that they were less likely to be speaking directly to individual devices. Finally, participants in the central condition more frequently shifted gaze from device to device while engaged in a task, suggesting that they were unconcerned with keeping eye contact with a specific device. Taken together, these results suggest that participants in the central condition behaved more like they were engaged with a single entity than those in the distributed condition.

Conclusion

The present study was intended to investigate how people speak to computational systems. Controlling whether users believed they were speaking to a single centralized system or to several separate devices, we found a centralized system was more efficient and easier to use than separate devices in several ways. Not surprisingly, the main difference was that users of the central system treated the system as a single entity whereas users of the separate devices treated the devices as independent entities. By relying on a single controller, users in the centralized condition were more likely to reuse conversational context than users in the distributed condition. Moreover, because they interacted with a single entity, users did not need to divide attention across several conversational partners.

What are some implications for Clark's joint activity view of conversation? It may initially seem misguided to apply this theory to human-computer interaction, for it was intended to deal with human-human interaction only. And afterall, computers and other devices are not *true* conversational partners because they are controlled by their users and cannot really engage in conversation. Nonetheless, Clark's theory was in fact predictive of behavior in this study, demonstrating that in this human-computer interaction context, many of the same assumptions about human-human interaction apply.

What are some implications for the design of future computing environments? First, for the sorts of tasks considered here, it is clear that a single controller is to be preferred over multiple devices. Thus, when designing a system that requires a user to coordinate information and activities among a set of distinct displays or information sources, it would be appropriate to provide the user a single point of contact with the overall system, as this would allow the user to establish an ongoing relationship with a single entity. Second, because maintaining context seems critical for efficiency (and possibly for ease of use as well), providing users with appropriate state information would likely encourage them to rely on established context. Third, because users tended to fix their gaze on individual devices in the multiple device condition, gaze cues (in addition to language cues) might be useful in helping the system determine level of engagement and to disambiguate referential statements, but cannot be relied on completely in the single controller case. Fourth, if assumptions about common ground can be manipulated by instructions, then the physical design of a system should be carefully considered. For example, putting several screens in the environment with the same physical size and characteristics might suggest multiple devices, whereas one large display and a few smaller displays might appear to be a single system with one point of contact and several output monitors. Fifth, giving devices obvious ways of collecting information from the room, such as a visible camera, allows users to understand what kind of common ground the system is likely to have. For instance, the camera may make users more likely to use gestures (e.g. pointing) to reference information. Finally, providing users with a single point of control need not have consequences for implementation; it might simply be enough to *tell users* to speak with a single device.

References

Argyle, M. & Cook, M. (1976). *Gaze and mutual gaze*. London: Cambridge University Press.

Clark, H. H. (1996). *Using language*. Cambridge England: Cambridge University Press.

Clark, H. H. & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies – why and how, in *Proceedings of the Workshop on Intelligent User Interfaces '93*.

Gould, J. D., Conti, J., & Hovanyecz, T. (1983). Composing letters with a simulated listening typewriter. *Communications of the ACM, 26(4),* 295-308.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 32*, 1-25.

Maglio, P. P., Matlock, T., Campbell, C. S., Zhai, S., & Smith, B. A. (2000). Gaze and speech in attentive user interfaces, in *Proceedings of the International Conference on Multimodal Interfaces 2000*.

Matlock, T. Campbell, C. S., Maglio, P. P., Zhai, S., & Smith, B. A. (2001) Designing feedback for an attentive office, in *Proceedings of Interact 2001*.

Maybury, M. T. (1997). Conversational multimedia interaction. In Y. Wilks, (Ed.) *Machine Conversations.* Kluwer Academic, Norwell, MA.

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

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

Perez-Quinones, M. A. & Sibert, J. L. (1996). A collaborative model feedback in human-computer interaction, in *Proceedings of the Conference on Human Factors in Computing Systems, CHI '96*.

Reeves, B. & Nass, C. (1996). *The media equation.* Cambridge, England: Cambridge University Press.

Tsukahara, W., & Ward, N. (2001). Responding to subtle, fleeting changes in the user's internal state, in *Proceedings of the Conference on Human Factors in Computing Systems, CHI 2001,* 77-84.

Yankelovich, N., Levow, G. A., & Marx, M. (1995). Designing SpeechActs: Issues in speech user interfaces, in *Proceedings of the Conference on Human Factors in Computing Systems, CHI '95*.