Towards Conversational Speech Recognition for a Wearable Computer Based Appointment Scheduling Agent

Benjamin A. Wong, Thad E. Starner, and R. Martin McGuire
College of Computing and GVU Center
Georgia Institute of Technology, Atlanta, GA, USA 30332-0280

We present an original study of current mobile appointment scheduling devices. Our intention is to create a conversational wearable computing interface for the task of appointment scheduling. We employ both survey questionnaires and timing tests of mock scheduling tasks. The study includes over 150 participants and times each person using his or her own scheduling device (e.g., a paper planner or personal digital assistant). Our tests show that current scheduling devices take a surprisingly long time to access and that our subjects often do not use the primary scheduling device claimed on the questionnaire. Slower devices (e.g., PDAs) are disproportionately abandoned in favor of devices with faster access times (e.g., scrap paper). Many subjects indicate that they use a faster device when mobile as a buffer until they can reconcile the data with their primary scheduling device.

The findings of this study motivated the design of two conversational speech systems for everyday–use wearable computers. The Calendar Navigator Agent provides extremely fast access to the user's calendar through a wearable computer with a head-up display. The user's verbal negotiation for a meeting time is monitored by the wearable which provides an appropriate calendar display based on the current conversation. The second system, now under development, attempts to minimize cognitive load by buffering and indexing appointment conversations for later processing by the user. Both systems use extreme restrictions to decrease speech recognition error rates, yet are designed to be socially graceful.

Categories and Subject Descriptors: []:

General Terms:
Additional Key Words and Phrases: appointment scheduling, context awareness, intelligent agents, speech recognition, wearable computing

1. INTRODUCTION

In science fiction, as well as science writing, conversational interfaces for computers have held a fascination. For example, Vannevar Bush in his 1945 paper "As We May Think" predicts that machines will one day be able to synthesize and tran-

This work funded in part by Starner's NSF Career Grant #0093291 and the NIDRR Wireless RERC.

scribe speech [Bush 1945]. Bush points to early prototypes of that era, the "voder" and "vocoder," which were demonstrated at the World Fair and at Bell Laboratories, respectively. "Wizard of Oz" prototypes of speech-controlled dictation in the 1970's and 80's demonstrated the feasibility of a "listening typewriter" [Gould et al. 1983]. Many concept videos in the 1990's, such as Apple's Knowledge Navigator, concentrated on anthropomorphic agents that can communicate via natural language. Also in the 1990's, systems began to recognize speech well enough for tasks such as ordering airline tickets [Kubala et al. 1994]. By 1994, research grade systems, such as variants of BBN's BYBLOS, could perform speech recognition with a 3000 word vocabulary on a 100MHz Intel 80486 processor with only 32M of RAM [Makhoul 1994]. With such technology being available for many years, what prevents the wide-spread adoption of conversational speech systems? For example, since the early 90's, most wearable computing vendors have emphasized speech recognition technology as a means for providing hands-free access to information. Today, even StrongArm based PDAs have sufficient power to perform speech recognition locally. Why aren't consumers talking to their PDAs and wearables instead of writing and typing?

2. PAPER OVERVIEW

In this paper, we examine conversational speech interfaces with respect to their potential use on wearable computers. We discuss reasons both for and against such interfaces and then argue that there may be applications and problem domains that are particularly well suited for speech interfaces.

We argue that a profitable avenue of research may be speech agents that "listen in" to the user's conversations and provide "just-in-time" information based on what the user is discussing, and we briefly discuss methods of addressing privacy issues with such agents. In particular, we choose the specific task of appointment scheduling to examine in detail.

In order to ground our research in conversational speech interfaces for appointment scheduling, we perform a survey of scheduling techniques currently used in the Georgia Tech population. We also examine, in detail, how 150 volunteers currently use memory, scrap paper, day planners, PDAs, and other devices to schedule mock appointments. Many subjects do not use the device that they claimed in the paper survey, and there is a high correlation between the frequency of disuse of a particular device and the amount of time required to access that device. We hypothesize that current appointment scheduling devices have deficiencies in ease of access and increased cognitive load if used in parallel with a conversation.

Based on these hypotheses, we design two functional speech agents, the Calendar Navigator Agent and Dialog Tabs that attempt to address these issues. Both of these agents are designed specifically for wearable computers with head-up displays. We discuss how speech recognition errors might be reduced through a combination of "push-to-talk" and limiting the phrases to those that are both easy to recognize and socially appropriate for the process of scheduling an appointment.

3. POTENTIAL BENEFITS OF SPEECH SYSTEMS FOR WEARABLES

While the authors are primarily interested in wearable computers, many of the arguments made in the literature for desktop-based conversational systems apply. Wearable computers, while not used in the same manner as desktop computers, may be used for many of the same applications as desktop computers. The situation is similar to when desktop computers began to be used in the place of mainframes. While many of the tasks were similar between the mainframe and the desktop (e.g. computing spreadsheets, editing text, etc.), the interfaces evolved to take advantage of the fact that one person probably controlled the computer and much of the computation could be dedicated to pleasing that user. With a wearable computer, a conversational system may assume technical capabilities similar to a desktop but can also take advantage of the mobility of the wearable, its physical closeness to the user, and the increased sense of privacy and control.

In 1983, Gould published "Composing Letters with a Simulated Typewriter" [Gould et al. 1983]. In this study, Gould uses a fast typist to simulate speech recognition systems with various capabilities such as isolated versus continuous recognition and 1000 versus 5000 versus unlimited vocabularies. He discovered that even the most limited of these systems was comparable in speed to handwriting for novice users dictating correspondence. In fact, for the continuous, unlimited vocabulary simulation, composition rate was significantly faster than writing (12.7 words per minute on average versus 7.8 wpm). Significantly, novice users adapted to the system quickly, even though the experimenters went to considerable trouble to make the system behave as machine-like as possible, including emulating a crude editing system. For users experienced with dictating, the simulated system was slower than dictating to a secretary or to a dictating machine (15.8 wpm for the fastest method, continuous unlimited vocabulary, versus 24.8 wpm for the dictating machine and 30.1 wpm for dictating to a secretary). In both experiments, the size of the simulation's vocabulary correlated with how well the system was liked. However, Gould warns that experienced dictators were less enthusiastic than novices for the simulated systems, and the experienced dictators were much more sensitive to failures of the vocabulary, even when 91% of the words were "recognized."

Gould's studies indicate speech systems may be particularly important for novice users. Speech input may seem more intuitive and may help occasional users of a system obtain the desired functionality at a speed faster than if they had to learn an alternative device, such as a keyboard.

While Gould's dictation simulation showed that speech recognition would be slower than other forms of dictation, some studies show that speech could significantly speed interaction in specific domains [Martin 1989; Rudnicky 1993]. With Schmandt's XSpeak system [Schmandt 1994], expert mouse users thought that XSpeak's speech-driven system for selecting a window on a graphical user interface (GUI) was faster than using a mouse, even though it was slightly slower. Thus, for certain applications, speech might be desirable because it is faster, or at least perceived faster, than other forms of input, even with expert users.

While intuitiveness and speed in limited domains can be sufficient reasons for using speech systems, Cohen and Oviatt provide a detailed list of when speech may be advantageous in "The Role of Voice Input for Human-Machine Communication"

[Cohen and Oviatt 1995]:

(1) When the user's hands or eyes are busy.
(2) When only a limited keyboard and/or screen is available.
(3) When the user is disabled.
(4) When pronunciation is the subject matter of computer use.
(5) When natural language interaction is preferred.

Four of these five situations address wearable computers directly. The first two most obviously apply to wearable computers. In past studies and products, wearable computers are designed to assist with some task in the physical world where the user's hands are not free [Najjar et al. 1997; Smailagic and Siewiorek 1994; Ockerman et al. 1997; Stein et al. 1998]. In fact, several authors have argued that one of the main benefits of wearables is that they are designed for secondary tasks in support of the user's primary task. [Rhodes 2000; Starner 1999]. Wearables are often adapted to use as little of the user's resources (e.g. eyes, hands, mind, etc.) as necessary to leave them available for interacting with the scenario at hand. Often such a scenario includes some sort of mobility. Thus, wearables can not use desktop keyboards or screens.

In addition, wearable computers have a long history of being used to support people with disabilities [Collins et al. 1977; Ross and Blasch 2000]. In fact, some of the first wearable systems were designed as lip-reading aides [Upton 1968]. Finally, this paper will argue that wearable speech agents might be useful for informational support during everyday conversations, addressing Cohen and Oviatt's last situation.

While much of the research discussed here is applicable to wearable computing, wearables are significantly different than their desktop counterparts. The next section will discuss current concerns with speech interfaces, adapting arguments in the literature to wearable computing as well as discussing the unique challenges of wearables.

4. SOME CONCERNS FOR SPEECH SYSTEMS ON WEARABLES

Recently, popular press articles and conference panel sessions have been critical of speech systems [Newman 2000; James 2002]. Such articles may be in response to consumer disappointment in commercial dictation systems as well as a reaction to earlier concept videos that portrayed anthropomorphic agents addressed through speech. However, conversational system researchers have written articles about the limitations of these systems and where they are most useful for many years [Schmandt 1994; Karat et al. 1999; Yankelovich et al. 1995; Oviatt 1999; Danis et al. 1994]. Shneiderman provides a brief overview of the issues in his "Limits of Speech Recognition" [Shneiderman 2000]

4.1 Error

One of the most important limitations is speech recognition errors. Both Gould's subjects and current users of speech products are often stymied by errors. While recognition systems will continue to improve, some error rate must be expected and designed for. A conversational speech application may allow correction of

these errors directly (e.g. through keyboards or an oral command language), repair the errors as part of a more natural conversational exchange ("I'm sorry, I said Wednesday the fifth."), use of redundant or contextual information (e.g. an utterance recognized as "September 31" might be corrected to "September 30" since there are only 30 days in September), or the application may be designed such that errors do not greatly affect functionality. The error correction and repair mechanisms must be chosen carefully for the application and expected user. For example, Karat et al. found that novice users of speech dictation systems could get trapped in cascades of errors as the commands used to correct the error were, themselves, misrecognized. Meanwhile, expert users simply employed the keyboard to correct the errors.

Unfortunately, the more subtle methods for error discovery and correction mentioned above will probably be unsuitable for wearable computers for some time. Mobility significantly confounds speech recognition, resulting in higher error rates, and restricts the types of devices and methods that may be used for error correction.

One example of this difficulty is the Lombard effect. In the presence of noise people speak differently, with increased amplitude, reduced word rate, and clearer articulation [Junqua 1993]. This effect seems to be a natural adaptation to a noisy environment. However, today's speech recognizers are not trained for the Lombard effect, which impacts performance. In addition, commercial speech recognizers have not sufficiently addressed the varying noise situations that occur during mobile speech, further reducing recognition accuracy. For example, bursty street traffic noise and microphone noise due to wind can significantly impact a recognition system through insertion errors. Finally, a mobile speech interface may have to perform well while the user is happy, sad, excited, nervous, etc. The user's mood may significantly affect the speech recognition error rate.

4.2 Social gracefulness and privacy

Another potential pitfall for a conversational wearable interface is social gracefulness. While cellular phones and earbuds now allow users to communicate by seemingly talking into air, a voice control interface for selling stocks such as shown during IBM's St. Mark's Square advertisement ("Down. Down. Over. Sell! Sell!") would still be considered odd if encountered in a park today. Even if such interfaces become commonplace, which is possible, the question of privacy is raised, both for the user and for people who might converse with the user. Speech is socially interruptive and hard to ignore. Thus, the user's discussions with the wearable would be eavesdropped by bystanders even if that was not their intention. On the other hand, bystanders may be reticent to talk to the user because they believe the conversation must be being recorded by the user's microphone.

4.3 Speed and Access

Speech is by its nature linear. It is difficult to search a speech waveform for content. It is also difficult to speed the playback speed above 3× of a normal speaking rate. In addition, speech is often transient and not retained.

4.4 Cognitive Resources

Shneiderman observes that speech interfaces may impose an additional cognitive load on the user that may interfere with the task at hand if that task requires forming new memories [Shneiderman 2000; Karl et al. 1993]. The cognitive science and human factors literature supports this hypothesis [Schacter 2001; Wickens 1984; Blackwood 1997]. In particular, brain imaging studies using functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) have shown that activity, or lack of activity, in the lower left frontal lobe while subjects are trying to remember (encode) a word list for later retrieval correlates with how well the subject will remember or forget those words, respectively [Wagner et al. 1998].

In general, the left hemisphere is thought to be responsible for processing words, while the right hemisphere is thought to process pictures. The left frontal lobe plays an important role during speech production, and the lower portions of the left prefrontal cortex also help contribute to the "phonological loop." The phonological loop is the part of short term memory that allows temporary storage of a small amount of linguistic information, such as a new vocabulary word or telephone number. Several studies have shown that distractor tasks, such as indicating when a certain trigger word is spoken or which box is demarked on a computer screen, can affect verbal memory severely [Cermak and Wong 1999; Okuda et al. 1998]. However, while memory is affected during the encoding process, it is most likely not affected by similar distractor tasks during recall [Cermak and Wong 1999; Craik et al. 1996]. In addition, simple distractor tasks, such as repeating the word "the" or performing repetitive physical motions, have little effect [Marsh and Hicks 1998; Shallice et al. 1994]. Schacter states "Some of these same frontal regions have been implicaed previously in working memory – holding information on-line for brief time periods. Although we do not yet know how these laboratory findings relate to everyday absent-minded errors, it is tempting to speculate that some of the frontal regions ... are 'captured' by distracting activities that preoccupy us and contribute to failed prospective memory" [Schacter 2001].

Continuing with such speculation leads to potential implications for conversational interfaces for wearable computers. If the user needs to remember how to construct a command to address the wearable interface and speak the command, might this process interfere with the user's ability to recall later the details of the primary task she was performing? In other words, in certain circumstances, might the use of the wearable computer actually impede the primary task?

While each of the concerns expressed here will be a continuing source of research topics, there may be opportunities for conversational interfaces on wearable computers which minimize these concerns while providing significant benefits. The next section examines this possibility for one set of applications.

5. DETERMINING SPEECH INTERFACES TO PURSUE

Even though there is significant challenge to creating conversational interfaces for wearable computers, there is also significant opportunity. [Roy et al. 1997] describes three scenarios for "wearable audio computing": continuous audio capture and retrieval, communications management, and disability aids. While these scenarios were intended to include more styles of interfaces than just the conversational, they

are revealing in how well they correspond to Cohen and Oviatt's suggestions. An example of a disability aid would be a speech recognition and synthesis system combined with software maps and a GPS to allow blind users to navigate a city [Ross and Blasch 2000]. Another example would be the lipreading aid mentioned earlier [Upton 1968]. Examples of communications managers include conversational interfaces information sources such as rolodexes, weather, stocks, and e-mail. Several such systems have shown promise in the literature in the past couple decades [Yankelovich et al. 1995; Schmandt and Arons 1984], and limited commercial variants are being offered in cellular phone services.

However, of these three scenarios, wearables seem particularly well poised for applications of continuous audio capture and retrieval. Wearable computers provide a unique amount of access to the user's life. By mounting small cameras on a head-up display and combining microphones with earphones, wearable computers can hear as the user hears and see as she sees. PDAs and cellular phones, in comparison, usually reside in the user's pocket until they are used and do not have access to as much of the user's context.

Such continual access to the user's context could be crucial for a speech-based augmented memory agent to assist its user. For example, office workers spend 35–80% of their time in spoken conversation. High-end managers, who may be in need of the most assistance due to the high number of interruptions they receive, are generally at the top end of this scale. In fact, opportunistic communications may account for up to 93% of these managers' workdays [Whittaker et al. 1994]. With so much spoken communication, certainly something can be done with the data! Yet mobile conversational speech recognition is known to be a very hard problem.

Some past projects have avoided speech recognition and stored the audio directly, using other cues, such as pen strokes, location, or time of day, for indexing the audio [Stifelman et al. 1993; Stifelman 1996; Whittaker et al. 1994; Wilcox et al. 1997]. Roy et al provide an overview [Roy et al. 1997]. Such systems are are directed at situations when the amount of spoken information is overwhelming, such as attending a conference. In a conference environment, speech recognition of conversations would be highly difficult. However, might there be applications where the user is similarly overwhelmed but the scenario is constrained enough such that speech recognition might be used? This question led us to appointment scheduling.

5.1 Appointment Scheduling

Part of the motivation for this project is informal observations on the use of PDAs for scheduling appointments since the introduction of the Apple Newton in 1993. When a junior member of a community schedules an appointment with a senior member, the senior member often requests an e-mail providing the time, date, and brief topic of the appointment. This request happens even though the senior member obviously has access to a PDA. When the senior member is queried, the reason given for this behavior is that the schedule on the PDA may not be up-to-date and has to be reconciled with a schedule maintained on a desktop PC. However, there also seems to be some resistance to accessing the calendar stored on a PDA during conversation.

Our hypothesis is that the access time for a PDA (e.g. pulling the PDA out of a pocket or purse, opening it, booting it, and finding the right application) is a barrier

to use. In order to reveal more about the process of appointment scheduling, we present a survey of what tools subjects believe they use for appointment scheduling and a comparison to videotaped mock appointment scheduling performed by the same subjects. We also provide initial results on the access time for various appointment scheduling tools. Finally, we present two prototype interfaces that use recognition of conversational speech to attempt to improve the appointment scheduling process.

6. CALENDAR USER SURVEY

We performed a user study in the Georgia Tech Student Center, asking for passersby to volunteer as human subjects. The study lasted for three days and had around 150 participants. The survey had two parts: each subject answered a short questionnaire and was videotaped as she scheduled a few mock appointments. The procedure took less than fifteen minutes for each participant and was done immediately after volunteering.

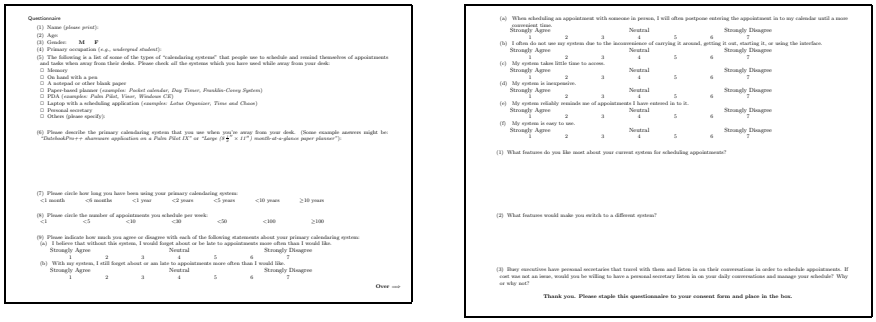
6.1 Questionnaire

We received 158 responses to our two–page questionnaire. The primary purpose of the questionnaire was to quantify which scheduling systems are in actual use and to gather qualitative impressions by the people who use those systems. We used eight Likert scale questions to rank each system in terms of effectiveness, ease of use, speed, and reliability. We ended the survey with open-ended questions to allow the subjects to fill in any gaps we might have missed.

6.1.1 *Questionnaire Setup.* The survey was performed in the lobby of the student center on the Georgia Tech campus and spanned three full days. As an incentive to participate, each subject was entered in a drawing for $128. Five different researchers over the three days recruited test subjects, distributed questionnaires and directed the subjects to the researcher who ran the timing test (see section 6.2). (The timing tests were always conducted by the same researcher). A small number of subjects answered the questionnaire *after* being tested in the timing test. Three researchers entered the survey results to computer and cross-checked their entries. The data and questionnaire is available electronically at `http://www.cc.gatech.edu/ccg/`. Figure 1 shows both sides of the actual questionnaire that was given to the participants.

6.1.2 *Questionnaire Results.* Because we performed the study in the student center of a technical institute, our demographics show a predominance of young, male students. (90% students, 88% age 18–25, 70% male)

In our sample population, the largest number of people (Figure 2) described their primary calendaring system, when mobile, as a "paper-based planner" (e.g., a pocket calendar). Note that Figure 2 does not include people who claimed more than one devices as a primary. Of the 158 people who responded, there were eighteen people who claimed multiple devices (Table I). Figures 3 through 10 are histograms that summarize the results of the Likert scale questions on the survey. Note that the histograms are aggregate values for all devices. The survey results are not different enough across scheduling devices to warrant separate charts. (The



Questionnaire

(1) Name (*please print*):
(2) Age:
(3) Gender: **M F**
(4) Primary occupation (*e.g., undergrad student*):
(5) The following is a list of some of the types of "calendaring systems" that people use to schedule and remind themselves of appointments and tasks when away from their desks. Please check all the systems which you have used while away from your desk:
☐ Memory
☐ On hand with a pen
☐ A notepad or other blank paper
☐ Paper-based planner (*examples: Pocket calendar, Day Timer, Franklin-Covey System*)
☐ PDA (*examples: Palm Pilot, Visor, Windows CE*)
☐ Laptop with a scheduling application (*examples: Lotus Organizer, Time and Chaos*)
☐ Personal secretary
☐ Others (*please specify*):

(6) Please describe the primary calendaring system that you use when you're away from your desk. (Some example answers might be: *"DatebookPro++ shareware application on a Palm Pilot IX"* or *"large (9 $\frac{1}{2}$" × 11") month-at-a-glance paper planner"*).

(7) Please circle how long you have been using your primary calendaring system:
<1 month <6 months <1 year <2 years <5 years <10 years ≥10 years

(8) Please circle the number of appointments you schedule per week:
<1 <5 <10 <30 <50 <100 ≥100

(9) Please indicate how much you agree or disagree with each of the following statements about your primary calendaring system:
(a) I believe that without this system, I would forget about or be late to appointments more often than I would like.
Strongly Agree Neutral Strongly Disagree
1 2 3 4 5 6 7
(b) With my system, I still forget about or am late to appointments more often than I would like.
Strongly Agree Neutral Strongly Disagree
1 2 3 4 5 6 7

Over ⟹

(a) When scheduling an appointment with someone in person, I will often postpone entering the appointment in to my calendar until a more convenient time.
Strongly Agree Neutral Strongly Disagree
1 2 3 4 5 6 7
(b) I often do not use my system due to the inconvenience of carrying it around, getting it out, starting it, or using the interface.
Strongly Agree Neutral Strongly Disagree
1 2 3 4 5 6 7
(c) My system takes little time to access.
Strongly Agree Neutral Strongly Disagree
1 2 3 4 5 6 7
(d) My system is inexpensive.
Strongly Agree Neutral Strongly Disagree
1 2 3 4 5 6 7
(e) My system reliably reminds me of appointments I have entered in to it.
Strongly Agree Neutral Strongly Disagree
1 2 3 4 5 6 7
(f) My system is easy to use.
Strongly Agree Neutral Strongly Disagree
1 2 3 4 5 6 7

(1) What features do you like most about your current system for scheduling appointments?

(2) What features would make you switch to a different system?

(3) Busy executives have personal secretaries that travel with them and listen in on their conversations in order to schedule appointments. If cost was not an issue, would you be willing to have a personal secretary listen in on your daily conversations and manage your schedule? Why or why not?

Thank you. Please staple this questionnaire to your consent form and place in the box.

Fig. 1. Questionnaire form used in survey of scheduling devices.

one possible exception is Figure 8, which would show a gap between the perceived expense of electronic and non-electronic scheduling devices).

The scale we used ranged from 1 (Strongly Agree) to 7 (Strongly Disagree). In Figure 3, "I believe that without this system I would forget about or be late to appointments more often than I would like," most subjects admitted some dependence on their system, with a mean of 2.9 (a neutral response was 4) and a standard deviation of 1.6. In Figure 4, "With my system, I still forget about or am late to appointments more often than I would like," subjects thought their systems were only somewhat effective, with a mean of 4.7 and a standard deviation of 1.7. Figure 5 shows the most flat and centered distribution over possible responses. The statement "When scheduling an appointment with someone in person, I will often postpone entering the appointment into my calendar until a more convenient time" resulted in an average score of 3.9 with a standard deviation of 1.9. In general, Figure 6, "I often do not use my system due to the inconvenience of carrying it around, getting it out, starting it, or using the interface," shows that subjects thought their systems were somewhat convenient to carry and access, with mean 5.0 and standard deviation 1.9. Subjects felt rather strongly that their systems were fast to access as shown by the 2.2 mean and 1.5 standard deviation in Figure 7, "My system takes little time to access." Figure 8 demonstrates that very many subjects felt strongly that their system was inexpensive, with a mean response of 2.0 and a standard deviation of 1.6 to the statement "My system is inexpensive." Figure 9, "My system reliably reminds me of appointments I have entered into it," with mean of 3.0 and variance of 1.7, indicates that subjects felt that, while their systems were good at providing appropriate alerts, there is room for improvement. In the survey, subjects felt most strongly, and with least variance, that their particular system was easy to use, with a mean of 1.8 and standard deviation of 1.2 in response to the statement "My system is easy to use," as shown in Figure 10.

6.1.3 *Questionnaire Discussion.* Most of the results from this part of the survey were not surprising; people are inclined to indicate that their devices are appropriate, sufficient, and somewhat necessary for reminding them of appointments. The questions related to overall effectiveness had moderately positive scores. Questions related to speed of access and ease of use were strongly positive and consistent. Ironically, these attributes which people felt most strongly about were not borne

No.	*Devices*
6	Memory + Paper
2	Memory + Planner
2	Paper + Planner
2	Paper + PDA
1	Memory + PDA
1	Paper + Skin
1	Planner + PDA
1	Memory + Paper + Planner
1	Memory + Paper + Skin
1	Memory + Other

Table I. Number of people who claimed more than one device as primary for mobile appointment scheduling

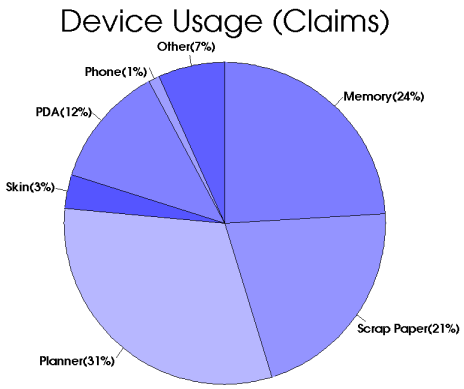


Fig. 2. Devices claimed as primary for mobile appointment scheduling. Paper planner, memory, and scrap paper combined account for 75% of the data at 31, 24, and 21 percent, respectively.

out by the observations made in the mock scheduling tasks (see section 6.2).

One possible explanation for this is that the answers for the devices with the largest number of respondants (planner, memory, and scrap paper) are overwhelming the contributions of less represented devices in the histograms. Surprisingly, the type of scheduling device seemed to have little or no effect on the answers to the survey: people claim to like their planners, PDAs, and scratch paper equally well. The only statement which elucidated a response attributable to device type was, "My system is inexpensive" which showed electronic devices as being more expensive.

The results for the statement "When scheduling an appointment with someone in person, I will often postpone entering the appointment into my calendar until a more convenient time," (Figure 5) are unusual. This was the only statement that resulted in no strong leaning one way or another. This may be an indication that people suspect that they do put off using their calendaring device, which would imply that they buffer the appointment somewhere until then. Or the neutral result could simply be an indication that our question was too vaguely worded and our

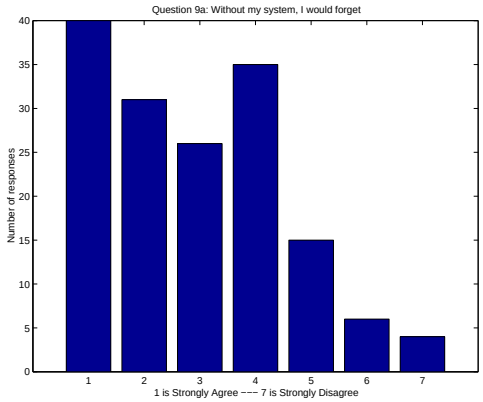


Fig. 3. "I believe that without this system, I would forget about or be late to appointments more often than I would like." (1 is strongly agree, 7 is strongly disagree). Mean: 2.9, std. dev.: 1.6.

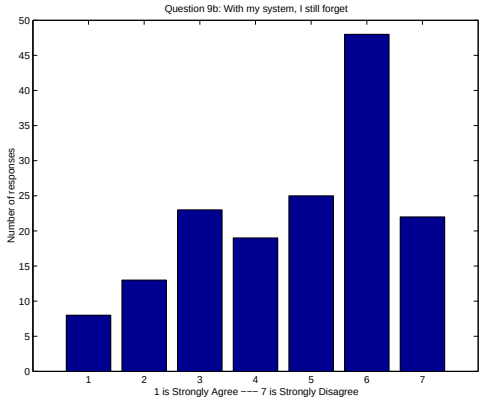


Fig. 4. "With my system, I still forget about or am late to appointments more often than I would like." Mean: 4.7, std.dev.: 1.7.

test subjects answered randomly.

One interesting result on the questionnaire was from one of our open-ended questions which was meant to probe the general sentiment to having an omnipresent personal secretary. This hypothetical situation is conceptually similar to the functionality a speech-driven wearable might provide. The results were clear: nearly two-to-one against having a personal secretary listening in on one's conversations throughout the day (*Yes*es: 53, *No*s: 96, *Maybe*s: 5). The most common reasons given were: wanting to be self-reliant, a desire for privacy, and lack of sufficient need. It is unclear how well this would apply to a wearable-computer providing a function similar to a personal secretary. For example, among the people saying they would not want a secretary because they prefer to do it themselves, were people who use PDAs and planners to assist them; implying that they see their devices as augmenting their abilities rather than replacing them.

Here are some comments typical of the people who answered negatively:

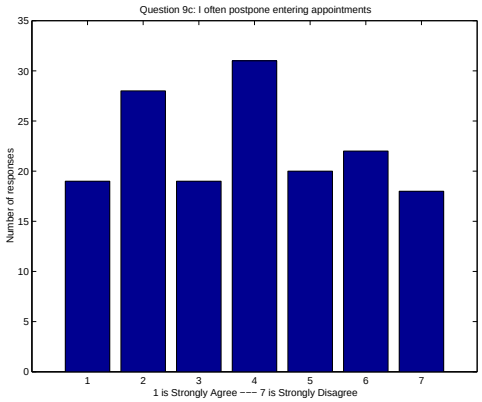


Fig. 5. "When scheduling an appointment with someone in person, I will often postpone entering the appointment into my calendar until a more convenient time." Mean: 3.9, std.dev.: 1.9.

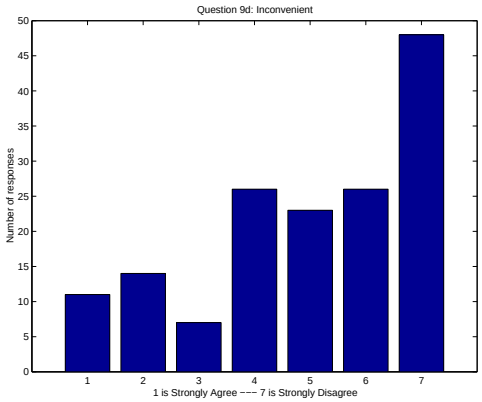


Fig. 6. "I often do not use my system due to the inconvenience of carrying it around, getting it out, starting it, or using the interface." Mean: 5.0, std.dev.: 1.9.

"It would feel like an invasion of my privacy."
"I want to manage my schedule by myself."
"Things like that should be one's own responsibility."
"I can manage, I like to be aware of my own activities so I remember them better."
"No, I'm not that busy that I feel it necessary to intrude upon my own privacy."
"No, I would prefer a machine to do the job."

6.1.4 *Future improvements with the survey method.* Most Likert scales place "disagree" on the left, and "agree" on the right. Our surveys did this in reverse and a few people had to cross out answers when they got confused. On the third day of the survey we corrected for this possible error by explicitly pointing out to each participant that the scale was "backwards". The data returned on the third day are statistically similar to that from the first and second days, so we believe

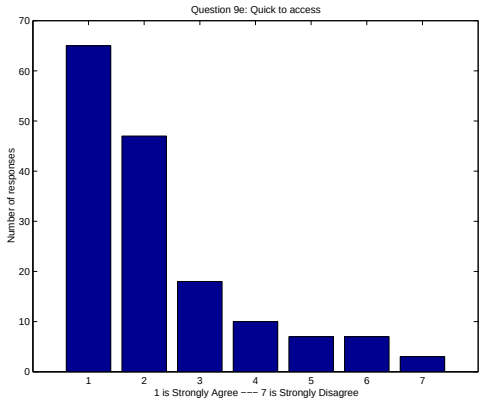


Fig. 7. "My system takes little time to access." Mean: 2.2, std.dev.: 1.5.

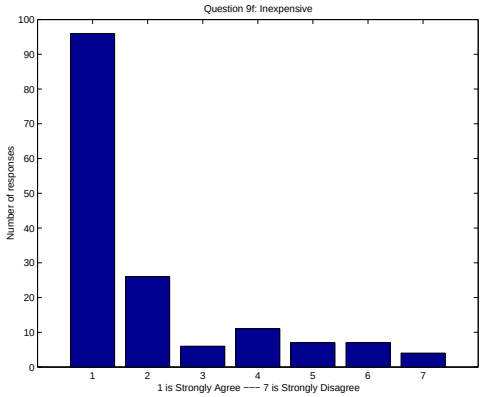


Fig. 8. "My system is inexpensive." Mean: 2.0, std.dev.: 1.6.

that the reversal of the scale did not significantly harm our results.

We allowed people to write in their scheduling system as an open-ended fill in the blank. This was both because we did not want to limit the devices to just what we knew of[1] and more importantly because some people use a *system* of multiple devices and we wanted to capture that, if possible. While we believe this ambiguity is more truthful to what people actually perceive as their primary systems, it sometimes added unnecessary complications, such as the exclusion of the 18 subjects who reported multiple devices in the pie chart (Figure 2) above.

6.2 Timing Test

We asked each participant to perform four scheduling tasks with their actual device and used a video recording to measure the amount of time the different phases of appointment scheduling took.

[1]None of our prior testing had turned up the possibility of a cell phone as a primary scheduler, for example.

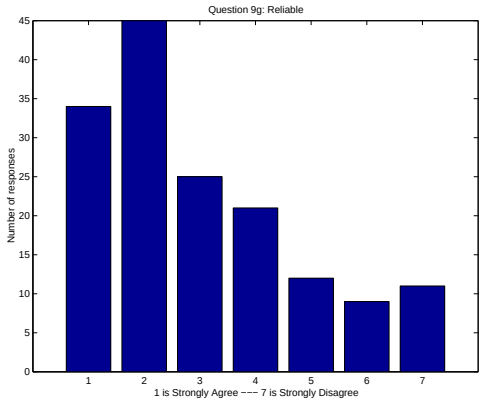


Fig. 9. "My system reliably reminds me of appointments I have entered into it." Mean: 3.0, std.dev.: 1.7.

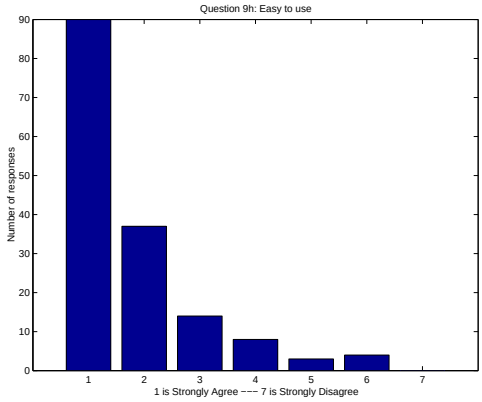


Fig. 10. "My system is easy to use." Mean: 1.8, std.dev.: 1.2.

The researcher used a script of four scheduling actions:

(1) Schedule an appointment for a date seven days in the future. *"Could we meet sometime next Monday?"*.

(2) Schedule an appointment for a date three months in the future. *"Could we schedule a time to meet in the second week of February?"*.

(3) Schedule an appointment for tomorrow. *"Could we schedule a time to meet tomorrow?"*.

(4) Reschedule appointment number two to another day. *"Could we reschedule our appointment in February from the 10th to the 11th?"*.

Although it is time consuming to extract timing data from video, the more obvious alternative of instrumenting each individual's scheduling device to record data would have sacrificed ecological validity and could not have covered as broad of a cross-section of the population. An additional, unintended benefit of the video was

that it showed a surprising discrepancy compared to what people claimed to use on their questionnaire. (See "Results" in section 6.2.2 below).

6.2.1 *Timing Test Setup.* In order to most accurately capture timing data from the video experiment, two cameras were used: one pointing forward, towards the test subject and a second standing on a tripod on the table, pointing down at where test subjects are likely to place their scheduling device while using it. (See Figure 11). The cameras were synchronized, and both also recorded audio.

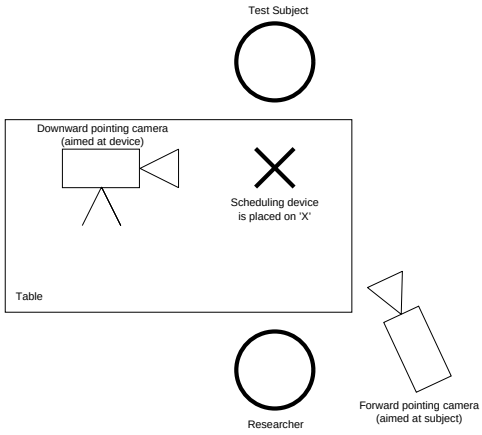


Fig. 11. Placement of cameras during appointment scheduling tasks captured on video.

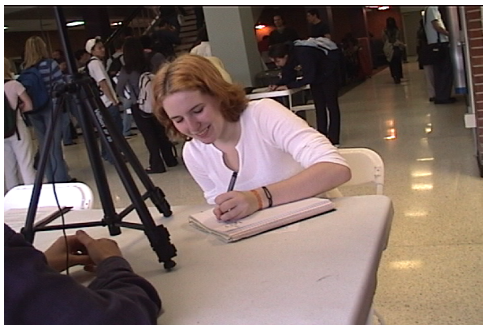


Fig. 12. A typical image from the forward-pointing video camera.

After all three days of the timing test was complete, several of the researchers transcribed the data from the video into a computer. This process of transcription took longer than gathering the video data in the first place. To lessen the chance of inconsistencies in the transcription, one of the researchers performed a few spot checks on the data gathered by the others and found his new numbers in perfect agreement. This may be because all the researchers were given clear instructions. For example "access time" is defined as the amount of time from when a test subject first starts looking for her scheduling device until the time when she has successfully

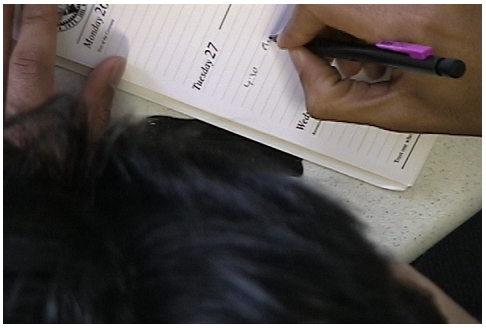


Fig. 13. A typical image from the down-pointing video camera.

retrieved it, navigated to the proper location in the device, and has her pen poised to enter a new appointment.

6.2.2 *Timing Test Results.* Access times were surprisingly high (see Figure 14). Also surprising was that the percentage of devices we saw (Figure 15) was very different than what was claimed on the questionnaire (Figure 2). This can also be seen (in Figure 16) by looking at the percentage of people (broken down by device) who did not use the device during the first task (or during any task). Finally, if we look at the ratio of actual usage to claimed usage (Figure 17) for devices with more than five data points in the survey, we can see that the devices are sorted in the same order as their access times: it appears that people are more likely to not use a device if that type of devices takes longer to access.[2]

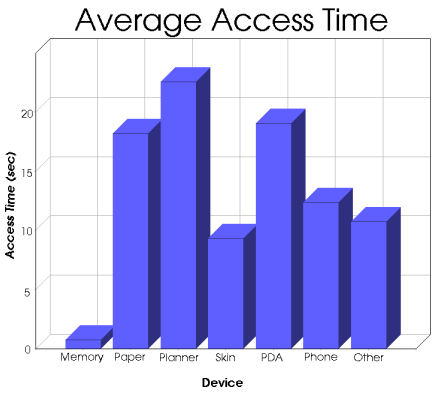


Fig. 14. Average time to retrieve a device and navigate to the location needed to enter an appointment. (Note that no attempt was made to measure the retrieval time of memory).

[2]Of course, it could also be the case that a device takes longer to access if people do not often take it out.

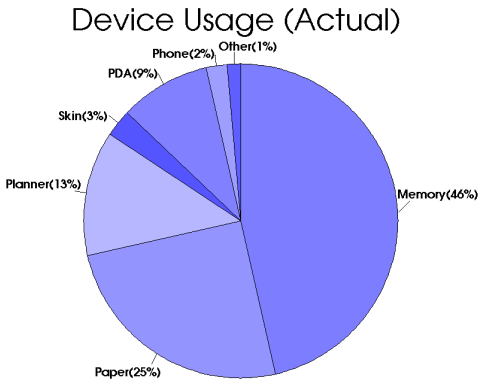


Fig. 15. Actual percentages of devices used as recorded on video

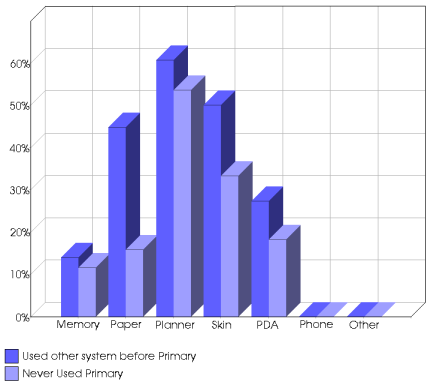


Fig. 16. Percentage of people who did not use the scheduling device they claimed as primary

Looking at Figure 17 more closely, it appears almost linear. How much of the ratio of disuse of a device can be explained as a linear relationship with that device's average access time? Using the data in Table 18, the correlation coefficient is determined. The result is a very strong negative correlation: -0.878, implying that 77% of the variance in disuse is explained by access time. An even stronger correlation can be calculated by making the reasonable assumption that "memory" actually takes a non-infinitesimal amount of time to access. The correlation becomes stronger and more linear until memory's access time is placed at (a probably unreasonable) 16 seconds: at that point the correlation coefficient reaches -0.992 and starts to taper off.

6.2.3 *Timing Test Discussion.* In the experiment above, appointment scheduling systems that are slower to access are disproportionately disused when compared to how many people claim to use them as their primary system. While the high correlation is striking, upon reflection, the result is not surprising. As early as 1968,

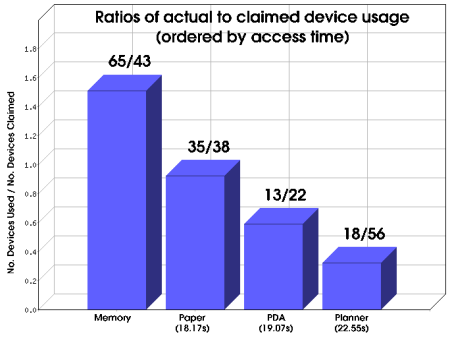


Fig. 17. Ratio of actual (video) to claimed (questionnaire) usage of devices ordered by access time. (Note that, because some participants claimed more than one device as "primary", the sum of the denominators is greater than the number of questionnaire respondents.

	Avg Access Time	*Ratio of Actual:Claimed Usage*
Memory	0?	65:31
Scrap Paper	18.38	36:24
PDA	20.42	13:18
Planner	22.55	17:50

Fig. 18. Table comparing average access time per device and the ratio of the number of people who used the device (on video) to the number who claimed to use the device (on the questionnaire).

Robert Miller argued

> . . . it will be easily demonstrated that many inquiries will not be made, and many potentially promising alternatives will not be examined by the human if he does not have conversational speeds—as defined in this report—available to him. But tasks will still be completed as indeed they have been in the past, without conversational interaction, and at least some of them will be completed more poorly by any criterion.

While Miller was describing the effects of waiting on a multitasking system, Rhodes continues this argument for subtasks that distract from a primary task [Rhodes 2000]. Here, negotiating an appropriate appointment time is the primary task, while retrieving the scheduling device is the subtask. Thus, with retrieval times in the tens of seconds, the subtask is a significant burden on the user and users are reluctant to use their device. This hypothesis corresponds with our informal observation on PDA users mentioned earlier in this paper. However,when the task becomes more complex, for instance with the scheduling of an appointment three months in advance, more subjects retrieve their device, as shown in Figure 16.

Another informal observation made both in the experiment and anecdotally in everyday life, is that very few subjects attempt to continue a conversation while entering an appointment into their device. This observation suggests that there is significant cognitive load in scheduling an appointment. Yet a quick observation of a busy city street will reveal people manipulating their calendars while walking. As Shneiderman pointed out in his analysis of uses of speech recognition, humans can

"speak and walk" easily but not "speak and think" [Shneiderman 2000]. A good hypothesis is that the manipulation of a appointment scheduler involves mechanisms that interfere with speech production. The brain imaging and short term memory studies mentioned earlier would seem to support this [Shallice et al. 1994; Schacter 2001]. For example, subjects may be using the phonological loop to temporarily store the potential appointment date and time while they perform the complicated physical actions of paging through their paper day planner or clicking menus on their PDA. Given these two tasks, the subjects have very little attention left to dedicate to conversation. The enforced social break in conversation or lack of attention forced by appointment scheduling (or possibly in even note-taking in general) may be another reason why owners of appointment scheduling devices are reticent to retrieve them unless necessary.

Interestingly, instead of using their primary device for simple appointments (e.g. meeting next week), subjects often used a lighter weight method of scheduling such as using their memory or writing on their skin or a piece of scratch paper. In some cases, the subject asked the experimenter to send e-mail so as to respond to the appointment request later. In analyzing the video data, over 66% of the subjects delayed reconciling the appointment with their primary scheduling system. When queried about this behavior, many subjects indicated that they were buffering the appointments until later. However, when a more complicated appointment was made (e.g. meeting three months in the future), more of the subjects would retrieve their claimed device to answer the appointment request. Perhaps, then, buffering behavior is a compensation for the longer access times or cognitive loads of the more complicated but complete scheduling systems.

The observations made here present a leading question: Can a device be made that supports appointment scheduling with low access time and/or low cognitive load? In the sections below, we will begin to explore this concept with two prototypes.

6.2.4 *Extensions to survey and timing test.* While the results presented above are striking and correspond with both the HCI and cognitive science literature, the subject population sampled was highly biased towards young, male students. Repeating the survey and experiment in an environment that is more representative of the population desired would be beneficial. Since our main purpose is to study the conversational interfaces for electronic appointment scheduling aids, directing our surveys towards populations with a heavier use of appointment scheduling devices seems appropriate. The travelers at Atlanta's Hartsfield Airport fit this description, and are, not coincidentally, also the market segment targeted by PDA manufacturers.

7. SCHEDULING PROTOTYPES

In the prior section we discussed possible interpretations of the appointment scheduling survey we performed. In this section we will describe two prototypes that begin to explore how to improve appointment scheduling devices.

To recap, the subjects in our survey often did not use what they consider their primary mobile scheduling system when confronted with an actual scheduling task. Instead, the subjects used lighter weight tools, often as temporary buffers for their

"real" scheduler.

These scheduling buffers postpone the task until later. Postponement may be beneficial as it reduces peak cognitive load by distributing the cognitive load over time. In addition, postponement allows appointments to be scheduled in a batch, which reduces the total set-up and tear down costs of using a scheduling device. Finally, some subjects stated that their upcoming appointments might conflict but that these appointments were not scheduled yet. In such a situation, postponement is necessary to avoid a scheduling conflict.

This research suggests two directions to pursue: what would happen if we created a lighter weight (faster to access) scheduler? Further, is it possible to explicitly employ the tactic of delaying cognitive load to create a better interface?

7.1 Calendar Navigator Agent

In general, appointments can not be scheduled without a dialogue between the participants making the appointment. With this in mind, no system could be more lightweight for the user than one that takes this dialogue, and nothing else, as input. One could imagine this ideal situation as a personal assistant who monitors the conversation. Without prompting from the user, the assistant listens to the normal scheduling dialogue and extracts the information necessary to record the appointment. Such an agent has little impact on the flow of the conversation except to interrupt if the proposed meeting time conflicts with a current appointment or to confirm the appointment was recorded correctly after the conversation is completed.

Wealthy executives and individuals sometimes have a butler or personal assistant that performs such services. In fact, the human assistant may even know and control the executive's schedule so that appointments can be made without the executive's active participation in the process. However, such systems are the subject for another paper. Here, we consider the assistant's role during a verbal negotiation. Specifically, we consider if a computerized version of such an assistant could be created such that the benefits of such an assistant can be brought to a larger population.

In the situation where the appointment scheduling process is a negotiation between the executive and another party, the assistant can, in fact, be a liability. The most obvious liabilities are the wages the assistant makes and the potential lack of privacy. However, the assistant may also prove interruptive and even slow down the process compared to the situation where the executive maintains a mental image of her calendar. To be more concrete, if the assistant notices that the conversation is converging on the date of June 12 at 2pm and that the executive already has a conflicting appointment, the assistant must interrupt the conversation orally and inform the participants. In addition, the assistant might suggest a later time, say 4:30pm, for the meeting. Since speech is linear in nature [Schmandt 1994], this interruption will take time. If, on the other hand, the executive knew that she was busy at 2pm, she could suggest another alternative without interrupting the flow of the conversation. Similarly, if the assistant could someone provide the executive with a visual representation of her calendar with the right dates and times appearing as the conversation progressed, the interaction would proceed more rapidly and with little interruption.

Is it possible to create such an appointment scheduling system that could, in

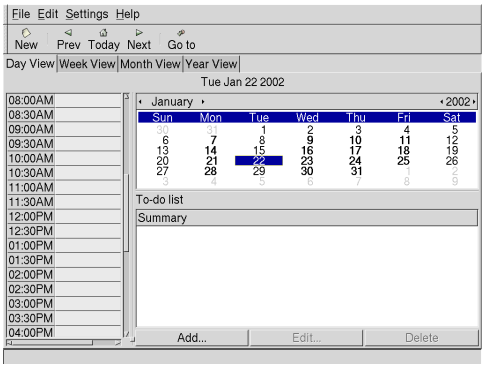


Fig. 19. Beginning Calendar Navigator Agent (CNA) interaction example. Subsequent figures captioned with spoken text that generated it. Bold text indicates speech marked for recognition by push–to–talk.

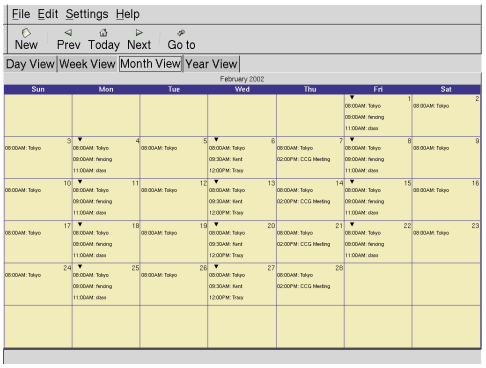


Fig. 20. CNA Input: "**February**? No, I'm going to be out of town for all of February."

theory, allow faster scheduling by changing the assistant's verbal interaction to a visual one? Could even a computerized version retain some of the potentially beneficial properties of such as system? The Calendar Navigator Agent (CNA) explores this idea.

7.1.1 *Implementation.* By outfitting the user with a wearable computer and a microphone, the system is able to listen in on the user's speech at all times. Still, limitations of current speech recognition technology make recognizing meaningful portions of casual conversation very difficult. However, by using push–to–talk techniques to allow the user to specify which parts of the conversation the computer should attend, the problem is simplified. The problem can be further simplified by restricting the grammar and vocabulary and informing the user of these limitations. In addition, the user can receive feedback and perform error correction through a calendar program displayed on the wearable's head-up, high resolution display.

Our implementation of the Calendar Navigator Agent begins with a freely available Unix calendaring system, Gnomecal (Fig. 19). As the user speaks, a button press instructs the system to begin recording. Recording stops with another press.

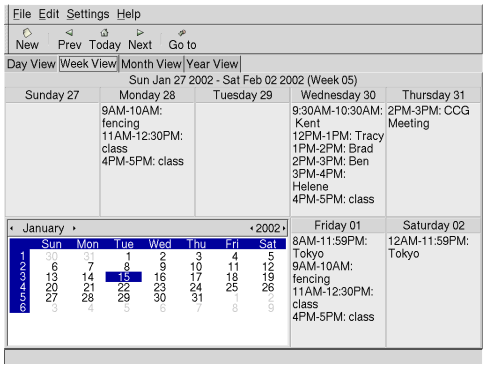


Fig. 21. CNA Input: "**Next week**? Let me see... Yup. How does **Tuesday** sound?"

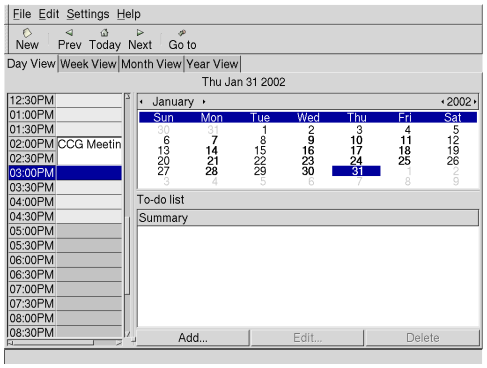


Fig. 22. CNA Input: "**Thursday**? I think so...hang on. **Next Thursday**? Yeah. I'm free at **3pm**."

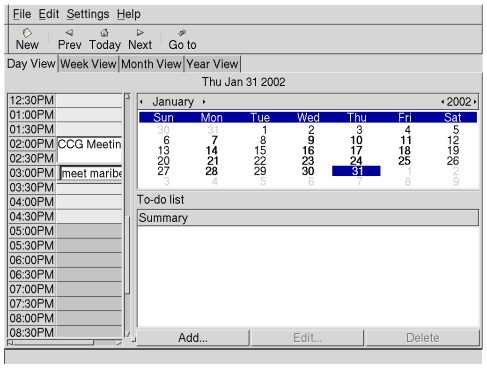


Fig. 23. CNA Input: "Okay, **see you then, Maribeth**."

The recorded speech fragment is then processed by IBM's ViaVoice engine using a very limited English vocabulary and grammar dealing mostly with dates and times. The output of ViaVoice is then passed into a program to generate the proper mouse and keyboard events to facilitate navigation and data entry within Gnomecal. By using push–to–talk techniques, we move the burden of discovering salient fragments of speech from the system to the user. Restricting the recognition grammar and vocabulary dramatically shortens processing time, increases accuracy, and encourages the user to employ more structured and grammatical speech when making appointments. Pursuing the concept further, the interface designer can provide users with socially graceful phrases which happen to be well perceived by the recognizer. Hidden Markov model based systems such as ViaVoice use context training and, as a result, tend to be more accurate with longer words and phrases. Thus, a phrase such as "Could we schedule an appointment for ..." could be used to start the CNA system. The trailing blank could be filled with "next week" or "next month" to determine what calendar page the CNA initially presents its user. Subsequent utterances and button presses can be constructed by the user to instruct the CNA as to what weeks, days, and hours to display. Through the rapid visual feedback, the user quickly sees when she is available and can schedule an appointment. The final acceptance of an appointment time is currently "See you then, ..." followed by the person's name. Note that the name might be recognized through speech from a small list of names or might be retrieved through other context sensors such as infrared identity badges.

Figures 19–23 illustrate a working prototype of the CNA used while filming a demonstration. While the system as shown is very limited, it can be executed in real-time on a wearable computer. As of yet, no speaker dependent training or additional context modeling has been exploited. Thus, we expect significant improvements in the performance of the system.

7.1.2 *Conversational interfaces that reinforce good conversational technique.* The last section described an interface where the user's speech had two purposes: informing an interlocutor and, simultaneously, cueing the interface. One would imagine that such speech might seem stilted or confusing. However, our initial experience with similar systems has been quite the opposite.

Our idea for socially graceful speech commands was influenced by a problem described in 1998 by the Boston Voice Users Group [DelPapa 1998]. One of the users who used a commercial speech package for his everyday work, noticed that it was very inconvenient and socially ungraceful to disengage the system when an unannounced guest visited his office. He had to say "Go to sleep" to his speech recognition system to stop recognition, turn to his visitor, say "Just a second", and then remove his headset and earpiece [DelPapa 1998]. By changing the stop recognition command on his speech recognition system to "Just a second," he performed two functions a once: disengaging his recognizer and informing his guest that he needed time before starting a conversation to remove his headset. In many senses, the user did not loose any flexibility of his system since "go to sleep" and "just a second" are probably equally unlikely to appear in his technical prose.

Another observation came from an initial experiment with the Remembrance Agent [Rhodes and Starner 1996], a system that continually examines the user's

typing and searches her system for similar text that may be useful to her task. The first author of this paper wore a noise cancelling microphone and a wearable computer for six months which would attempt to transcribe his everyday speech and enter it into a text buffer. The Remembrance Agent then used this text buffer for its searches. While the experiment is still being refined, the experience was somewhat surprising from the point of view of conversational partners. During technology open houses and casual interactions, the second author (the first author's advisor) repeatedly received compliments about how well spoken the student was. In addition, the advisor noticed that the student employed very good conversational technique, verbally confirming appointments and tasks. Of course, the student was actually repeating the information not as good conversational technique but instead for the benefit of speech recognizer so that it would transcribe the information into his text buffer for later use!

Since speech recognition engines may be many years before they can detect and transcribe conversational speech made by mobile interlocutors who are not wearing noise canceling microphones, such echoing behavior may be the most efficient way of cueing the CNA. If the user's conversational partner suggests a date, the user can respond with an appropriate phrase such as "Yes, I believe I can meet you Tuesday at 2pm." This utterance cues the CNA as well as informs the interlocutor that the user understood the suggestion correctly. Whenever specifics are discussed, such "echoing" in conversation is encouraged by professionals, from the military to professional speech trainers. Thus, the CNA interface may actually reinforce good conversational skills.

7.1.3 *Privacy.* Due to its design, the CNA protects the privacy of the user's conversational partners on several levels. One of the largest concerns bystanders have with wearable computers is that they are being recorded [Strubb et al. 1998]. Audio recording is particularly sensitive topic. However, the noise canceling microphone worn by the CNA's user barely hears the speech of anyone except the user. Unless an interlocutor is shouting, the speech is almost indecipherable without careful preprocessing. In addition, the interlocutor's speech is discarded by the user's speech recognition engine as unrecognizable. In fact, even the user's speech is discarded after the cues for navigating the CNA are extracted. Thus, casual conversational partners of a CNA user are protected from audio recording on the physical level, on a systems level, and on an application level. While a nefarious wearable user could subvert all of these by adding a hidden ambient-level microphone to his system, we have found that most bystanders' privacy concerns are addressed simply by demonstrating the comparative audio signal levels of the user and the interlocutor.

7.1.4 *Challenges with the CNA.* With extensive development and user skill, the CNA might approach the ideal of low access time and social gracefulness, especially if the user's display was integrated into her eyeglasses as demonstrated by the MicroOptical Corporation [Spitzer et al. 1997]. However, the CNA places significant additional burdens on the user. The cognitive overhead of formatting speech for the recognizer, as well as marking the start and stop times of an utterance for recording, may be prohibitively high. In addition, even with the harsh restrictions,

Fig. 24. To limit distractions Dialog Tabs take up very little visual area (they are the vertical bars on the right 1% of the screen).

speech recognition is plagued with errors. If the CNA misses a cue, for example, not updating the screen when the user says "Let me see if I'm available next week," the user must either rephrase the same request again in a socially appropriate way or apologize to her partner and explicitly address the system, via speech or keyboard, to correct the error. In the next section, we discuss a system that is designed to minimize cognitive load and miscueing due to speech recognition errors.

7.2 Dialog Tabs

While access time does seem to be correlated with the probability that a device will not be used, it is clearly not the only factor. Our survey suggests that the subjects may use a buffer (such as their memory or scrap paper) when mobile to delay the burden of dealing with their primary scheduling device. As mentioned previously, postponement may help reduce peak cognitive load, enable entering appointments in batch, or delay commitment until potentially conflicting appointment dates are fixed.

We are currently designing a system called "Dialog Tabs" to explore how speech recognition and audio capture might be used to aid postponing the processing of appointments. The system is designed for low attentional demands during conversation as well as fast access for when the user wants to process the appointment.

7.2.1 *Implementation and relation to current work.* Unlike the CNA, Dialog Tabs do not require a push–to–talk interface. Rather, Dialog Tabs use an open microphone which continuously monitors all utterances through-out the day. When the user speaks an appointment time during normal conversation, a form of dialog box, called a "dialog tab", is displayed. It is non-modal and appears as a small (two pixel wide) tab on the right side of the heads-up display (fig. 24).

These tabs "contain" a audio waveform of the speech that is suspected to contain an appointment scheduling event as well as information that could be parsed by the speech recognizer. As new possible appointment scheduling events occur, the tabs are stacked in order of arrival. The most recent tab is longest, covering the top half of the right edge of the screen. The next most recent tab is second longest, covering the next quarter of the edge. The third most recent covers one eighth of the screen, and so on in a series for as many tabs are displayed.

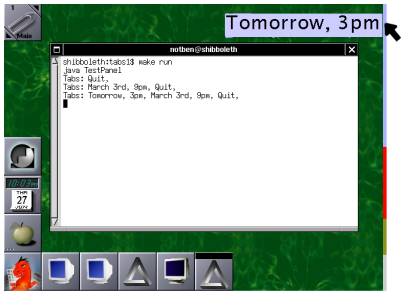


Fig. 25. Dialog Tabs are arranged along the edge of the screen to allow for extremely rapid interaction. Hovering the mouse over a tab reveals the information contained within.

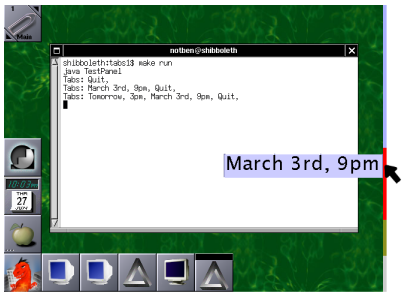


Fig. 26. When a new tabs appears, it is place on the top. Older tabs are stacked below, with geometrically decreasing size

The Dialog Tabs require little screen real estate and are designed to be minimally distracting even when they appear on a head-up display during conversation. During the day, dialog tabs may queue up on the side of the screen, but the user does not need to process them until she has spare time or the inclination to process her appointments in batch. The tabs provide a constant reminder that the user needs to process her appointments. Thus, the user can postpone processing the appointment events without fear of forgetting them.

However, Dialog Tabs are also being developed so that they can be processed quickly (e.g. as the user walks to his next appointment). Dialog Tabs takes advantage of Fitts's law by placing the tabs on the edge of the screen rather than the center [Walker and Smelcer 1990]; if the user wants to click on a tab it should be easy to do so even in poor motor control situations such as while walking [Lyons and Starner 2001]. Hovering over a tab displays the date discussed in the user's conversation (Figures 25, 26) as parsed by the speech recognition engine. The dialog tab may have been produced in error, in which case it can be dismissed by a right click. A left click of the mouse immediately accepts the appointment contained in the dialog tab and places the appointment in the user's calendar. However, if the dialog tab is a valid appointment but the date was recognized incorrectly, a middle click will open the calendar to the suspected date and time and allow the user to fill-in or correct any fields before committing the appointment to the appointment calendar. If the user cannot remember the specifics of the appointment, the audio

from the event can be played back for the user.

To date the preliminary interface for Dialog Tabs has been produced as shown in the Figures, but the speech recognition engine has not been integrated, Again, speech recognition errors will prove onerous. However, recent research by Whittaker on a project called SCANmail brings hope [Whittaker et al. 2002]. SCANmail applies speech recognition to the problem of voicemail. Each message is stored as a audio waveform with its caller identity and time. In addition, the speech recognition engine attempts to parse the most important and recognizable pieces of the message. In most cases, the speech engine parses strings of telephone numbers, and SCANmail displays these numbers as a summary to the user. Often, such telephone numbers are the most critical part of the message, and users of voice mail systems have to repeatedly play their voice mail to get the appropriate phone number. SCANmail allows users to click on the summary words or numbers to play the corresponding part of the saved audio track which corresponds to that utterance. In this way, users can confirm that the speech recognizer had no errors in parsing the phone number. SCANmail uses speech recognition, even with poor general accuracy and an unconstrained vocabulary and grammar, to allow more rapid indexing and manipulation of voicemail messages. Dialog tabs will attempt to provide a similar index for potential appointments made during the user's day.

Note, however, that speech recognition is not the only way Dialog Tabs could recognize that an appointment event was occurring. To assist the system, the user could press a key or otherwise signal that a dialog tab should be made. Such explicit signaling will probably be the basis of the first deployment. However, in preparation for some form of speech recognition, the Dialog Tabs system has been made such that false positives can be quickly dismissed by the user. Note that an interface similar to Dialog Tabs may prove useful more generally for quick creation of audio notes to spur the user's later memory. In the past, a popular concept for a wearable computer suggested by novice users is one which would remember the past 30 seconds of a conversation when the user pressed a button. However, few users have suggested methods of accessing that data or indexing its contents for later retrieval. Dialog Tabs begins to address this similar issue in a constrained domain.

7.2.2 *Potential objections to Dialog Tabs.* Besides speech recognition errors, which can hopefully be addressed in a manner similar to SCANmail with active user support, there are two major potential objections to the Dialog Tab interface. First, will users remember enough of their appointment scheduling conversations at the end of the day in order to use the interface? Journal studies by Wagenaar [Wagenarr 1985] and more recent studies comparing recall and recognition [Schacter 2001] would seem to indicate affirmatively. Even when test subjects could not independently recall an event, providing more information about the event seems to allow subjects to recall progressively more independent facts about the event [Wagenarr 1985]. The audio stored by the Dialog Tabs, if not the parsed appointment dates, should be sufficient to cue users' memories.

Another, potentially more serious issue is how will a user of Dialog Tabs negotiate scheduling conflicts? One option is to have the dialog tab become more visible and show a warning if it thinks that a scheduling conflict is about to happen.

However, such a system may be very unreliable due to speech recognition errors. Another option is simply to let the user deal with the conflict when she processes the Dialog Tab. While seemingly inefficient, the concept of tentative appointments is well accepted in the world of PDAs which require synchronizing with desktops before receiving the most up-to-date version of a schedule. In many senses, then, the potential negative results of the postponement that the Dialog Tabs enables is no worse than current systems and is currently accepted socially. Thus, the main benefit the Dialog Tabs may offer is the relative ease and low cognitive load of recording a reminder for the later entering of an appointment into the user's calendar. Additional features, such as conflict alerts, better speech recognition for automatic recognition of appointment events, and summaries of appointment event negotiations, while convenient, may not be necessary for initial deployment of the system.

8. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we have discussed some of the promise and challenges for conversational systems on wearable computers. We have chosen the task of appointment scheduling to examine in detail for developing a conversational agent interface and have performed a survey of current appointment scheduling systems to inform our research. The results of this survey, when combined with the results of a videotaped mock appointment scheduling task held in parallel, suggest that current appointment scheduling systems suffer from inconveniently long access times and possibly high cognitive load. Longer access times for a class of device correlated highly with the probability that a given device was not used by its owner during mock scheduling tasks. In the future, repeating the calendar survey and timing test in an area with a high concentration of busy professionals who may use electronic scheduling aids (for example, Atlanta's airport) would provide a convenient comparison with the study we performed at Georgia Tech's student center.

To explore the issues of access time and cognitive load in appointment scheduling systems, we have created two prototypes based on interacting with conversational speech: the Calendar Navigator Agent (CNA) and Dialog Tabs. The CNA monitors appointment scheduling conversations and displays calendar information based on the progression of that conversation. In order to constrain the speech recognition problem, the user exploits a variant of push-to-talk and maintains a grammar of socially appropriate but narrowly defined utterances. In the future, we wish to test the Calendar Navigator Agent against other appointment scheduling tools, such as paper-based day planners, PDAs, and human assistants, to determine the CNA's relative accessibility and effectiveness.

The Dialog Tabs attempt to encapsulate appointment scheduling conversations for later processing by the user. The Dialog Tabs provide a later reminder about a possible appointment without forcing the user to interact with a distracting calendar interface during a conversation. While the system is still under development, it shows promise as a more general reminder system for wearable computers. Eventually, we wish to test Dialog Tabs on mock conversations over the course of several days to determine its utility and usability.

ACKNOWLEDGMENTS

The authors would like to thank Amy Hurst, Brad Singletary, David Minnen, Helene Brashear, and Mel Eriksen for their help in this project. We also thank our hosts at the Wearable Computing Laboratory at the Swiss Federal Institute of Technology Zurich (ETH) for their resources and hospitality. Much of this paper was written on CharmIT Pro wearable computers. This material is based, in part, upon work supported by the National Science Foundation under Grant No. 0093291. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This publication is also supported, in part, by the Rehabilitation Engineering Research Center on Mobile Wireless Technologies for Persons with Disabilities, which is funded by the National Institute on Disability and Rehabilitation Research of the U.S. Department of Education under grant number H133E010804. The opinions contained in this publication are those of the grantee and do not necessarily reflect those of the U.S. Department of Education.

REFERENCES

Blackwood, W. 1997. *Tactical Display for Soldiers*. National Academy of Sciences, Washington, D.C.

Bush, V. 1945. As we may think. *Atlantic Monthly 76,* 1 (July), 101–108.

Cermak, L. and Wong, B. 1999. The effects of divided attention during encoding and retrieval on amnesic patients' memory performance. *Cortex* 35, 73–86.

Cohen, P. and Oviatt, S. 1995. The role of voice input for human-machine communication. In *Proceedings of the National Academy of Sciences*. Vol. 92. 9921–9927.

Collins, C., Scadden, L., and Alden, A. 1977. Mobility studies with a tactile imaging device. In *Fourth Conf. on Systems and Devices for the disabled*. Seattle, WA.

Craik, F., Govoni, R., Naveh-Benjamin, M., and Anderson, N. 1996. The effects of divided attention on encoding and retrieval processes in human memory. *J. of Experimental Psychology: General* 125, 159–180.

Danis, C., Comerford, L., Janke, E., Davies, K., DeVries, J., and Bertrand, A. 1994. Storywriter: A speech oriented editor. In *CHI*. Boston, MA, 277–278.

DelPapa, J. 1998. Personal communication. *Boston Voice Users Group*.

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

James, F. 2002. Panel: Getting real about speech: Overdue or overhyped. In *CHI*.

Junqua, J. 1993. The lombard reflex and its role on human listeners and automatic speech recognizer. *J. Acoustic. Soc. Amer. 93*, 510–524.

Karat, C., Halverson, C., Horn, D., and Karat, J. 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *CHI*. 568–572.

Karl, L., Pettey, M., and Shneiderman, B. 1993. Speech versus mouse commands for word processing applications: An empirical evaluation. *Intl. J. Man-Machine Studies 39,* 4, 667–687.

Kubala, F., Anastasakos, A., Makhoul, J., Nguyen, L., Schwartz, R., and Zavaliagkos, G. 1994. Comparative experiments on large vocabulary speech recognition. In *ICASSP*. Adelaide, Australia.

Lyons, K. and Starner, T. 2001. Mobile capture for wearable computer usability testing. In *Submitted to IEEE Intl. Symp. on Wearable Computers*. Zurich, Switzerland.

Makhoul, J. 1994. Personal communication. *BBN Speech Systems*.

Marsh, R. and Hicks, J. 1998. Event-based prospective memory and executive control of working memory. *J. of Experimental Psychology: Learning, Memory, and Cognition* 24, 336–349.

Martin, G. 1989. The utility of speech input in user-computer interfaces. *Intl. J. of Man-machine studies 30,* 4, 355–375.

Najjar, L., Thompson, C., and Ockerman, J. 1997. A wearable computer for quality assurance inspectors in a food processing plant. In *IEEE Intl. Symp. on Wearable Computers*. IEEE Computer Society.

Newman, D. 2000. Speech interfaces that require less human memory. *Speech Technology*.

Ockerman, J., Najjar, L., and Thompson, C. 1997. Wearable computers for performance support. In *IEEE Intl. Symp. on Wearable Computers*. IEEE Computer Society.

Okuda, J., Fujii, T., Yamadori, A., Kawashima, R., Tsukiura, T., Kukatsu, R., Suzukie, K., Ito, M., and Fukuda, H. 1998. Participation of the prefrontal cortices in prospective memory: Evidence from a PET study in humans. *Neuroscience Letters* 253, 127–130.

Oviatt, S. 1999. Ten myths of multimodal interaction. *Communications of the ACM 42,* 11, 74–81.

Rhodes, B. and Starner, T. 1996. Remembrance agent: A continuously running automated information retreival system. In *Proc. of Pract. App. of Intelligent Agents and Multi-Agent Tech. (PAAM)*. London.

Rhodes, B. J. 2000. Just-in-time information retrieval. Ph.D. thesis, MIT Media Laboratory, Cambridge, MA.

Ross, D. and Blasch, B. 2000. Wearable interfaces for orientation and wayfinding. In *ACM conference on Assistive Technologies*. 193–200.

Roy, D., Sawhney, N., Schmandt, C., and Pentland, A. 1997. Wearable audio computing: A survey of interaction techniques. Tech. rep., MIT Media Lab. April.

Rudnicky, A. 1993. Mode preference in a simple data-retrieval task. In *ARPA Human Language Technology Workshop*. Princeton, New Jersey.

Schacter, D. 2001. *The Seven Sins of Memory*. Houghton Mifflin, Boston.

Schmandt, C. 1994. *Voice Communication with Computers*. Van Nostrand Reinhold, New York.

Schmandt, C. and Arons, B. 1984. Phone slave: A conversational telephone messaging system. *IEEE Transactions on Consumer Electronics CE-30* 3 (August), 21–24.

Shallice, T., Fletcher, P., Frith, C., Grasby, P., Frackowiak, R., and Dolan, R. 1994. Brain regions associated with acquisition and retrieval of verbal episodic memory. *Nature 368*, 633–635.

Shneiderman, B. 2000. The limits of speech recognition. *Communications of the ACM 43,* 9 (September).

Smailagic, A. and Siewiorek, D. 1994. The CMU mobile computers: A new generation of computer systems. In *COMPCON '94*. IEEE Computer Society Press, 467–473.

Spitzer, M., Rensing, N., McClelland, R., and Aquilino, P. 1997. Eyeglass-based systems for wearable computing. In *IEEE Intl. Symp. on Wearable Computers*. IEEE Computer Society.

Starner, T. 1999. Wearable computing and context awareness. Ph.D. thesis, MIT Media Laboratory, Cambridge, MA.

Stein, R., Ferrero, S., Hetfield, M., Quinn, A., and Krichever, M. 1998. Development of a commercially successful wearable data collection system. In *IEEE Intl. Symp. on Wearable Computers*. IEEE Computer Society.

Stifelman, L. 1996. Augmenting real-world objects. In *CHI '96*.

Stifelman, L., Arons, B., Schmandt, C., and Hulteen, E. 1993. Voicenotes: A speech interface for a hand-held voice notetaker. In *CHI*. 179–186.

Strubb, H., Johnson, K., Allen, A., Bellotti, V., and Starner, T. 1998. Privacy, wearable computers, and recording technology. In *IEEE Intl. Symp. on Wearable Computers*. Pittsburg, PA, 150–154.

Upton, M. 1968. Wearable eyeglass speechreading aid. *American Annals of the Deaf 113*, 222–229.

Wagenarr, W. 1985. My memory: A study of autobiographical memory over six years. *Cognitive Psychology* 18, 225–252.

Wagner, A., Schacter, D., Rotte, M., Koutstall, W., Maril, A., Dale, A., Rosen, B., and Buckner, R. 1998. Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science* 281, 1188–1191.

Walker, N. and Smelcer, J. 1990. A comparison of selection time from walking and bar menus. In *Proceedings of CHI'90*. ACM, Addison-Wesley, 221–225.

Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamchick, G., and Rosenberg, A. 2002. Scanmail: a voicemail interface that makes speech browsable, readable and searchable. In *CHI*. ACM Press, New York, 275–282.

Whittaker, S., Hyland, P., and Wiley, M. 1994. Filochat: Handwritten notes provide access to recorded conversations. In *CHI*. ACM Press, New York, 271–276.

Wickens, C. 1984. *Varieties of Attention*. Academic Press, New York, Chapter Processing resources in attention.

Wilcox, L., Schilit, B., and Sawhney, N. 1997. Dynomite: A dynamically organized ink and audio notebook. In *CHI*. 186–193.

Yankelovich, N., Levow, G., and Marx, M. 1995. Designing SpeechActs: Issues in speech user interfaces. In *CHI*. 568–572.