Conversational Speech Recognition for Creating Intelligent Agents on Wearables

Benjamin A. Wong and Thad E. Starner
Future Computing Environments Group,
College of Computing and GVU Center
Georgia Institute of Technology, Atlanta, GA, USA 30332-0280

{bbb, thad}@cc.gatech.edu

Abstract

Paper-based planners and calendar programs run on personal digital assistants are common methods to remind users of tasks and appointments. Both require a certain amount of time to record an entry. Is it worthwhile (or even possible) to try to create an appointment scheduling system which takes significantly less time and effort? We argue that it is and describe one such system currently being built, its potential benefits, and some of the issues it must overcome. This system, called the Calendar Guardian Agent (CGA), runs on a wearable computer to assist in scheduling appointments based on captured speech from a user's everyday conversations.

1 Introduction

There are a variety of methods to remind people of tasks and appointments. These methods include paper-based planners, such as the Franklin Day Planner™, and calendar programs run on personal digital assistants (PDAs), such as the Palm Pilot™. Often these scheduling systems are used during a conversation in which a person is orally scheduling a meeting: "I'll meet you at 3pm tomorrow." However, using these current scheduling systems can incur a significant overhead. There are other methods of scheduling which do not require a large overhead in time, for example, using a human assistant or just trying to memorize an appointment. However those methods have considerable trade-offs in cost efficiency and reliability, respectively.

Wearable computing's capability for continuous interaction permits new techniques to facilitate everyday life and social interaction. One of our latest projects, the Calendar Guardian Agent (CGA), is a system to schedule appointments using an agent on a wearable computer which attends to the user's everyday conversations. Although we use techniques similar to a human assistant and memorization to reduce the overhead of scheduling, we expect the costs in terms of monetary expense and unreliability to be reasonable. Note that while portions of the CGA have been completed, the system described in this paper is still under development.

2 Motivation

2.1 Can scheduling be more efficient?

An interesting comparison to make for any scheduling system is to consider that system's convenience versus a hypothetical *human* assistant. For example, a personal secretary can be non-intrusive but always listening, ready to jot down another appointment or remind one of an upcoming event. Considering just the amount of time spent to operate a PDA or paper-based planner it becomes clear that it would be much more convenient to use a real personal assistant.

In an informal study of both paper-based planners and PDAs, we found that for appointments that take around 80 seconds to schedule, over three-fourths of the time was spent in the overhead of managing the scheduling systems, themselves. We simulated scheduling events by having five pairs of people orally arrange to meet at a specified date six months in the future. The initiator used a script and did not actually spend any time scheduling; the other used their ordinary scheduling system (PDA or paper-based planner). The average result was around 80 seconds. As a baseline we also simulated and timed scheduling events where each party assumed they had a personal assistant listening in, taking care of remembering the details. The average result of that test was around 20 seconds. To calculate a lower-bound for the overhead we found the difference in the time it took to schedule an appointment using a human assistant versus a traditional system. Our study was too crude to show any significant difference between PDAs and paper-based planners. Note that a human assistant achieves a lower overhead but greatly increases the monetary expense.

Once the burden of using a scheduling system rises above a certain threshold, a busy person might resort to memorizing appointments so that she can enter them later when scheduling can be her primary task. This happens despite the unreliability of human memory because of the low cost in time. In general, small increases in system delays can trigger large drops in usability [Rhodes, 2000, Shneiderman, 1998].

2.2 How overhead might be reduced

One of the key features that separates both a personal human assistant and memorization from a PDA or paper-based planner is the speed with which the "knowledge in the head" (to use Don Norman's term, [Norman, 1988]) is transfered to "knowledge in the world". Norman describes "knowledge in the head" as information known by the user (in this case, the appointment details the user knows). Whereas "knowledge in the world" is information that is stored in the environment (the physical artifact in the world that helps the user to remember). Both PDAs and paper-based planners require the user to explicitly transfer the knowledge from her head to the world. A personal assistant, on the other hand, can *listen in* on the conversation so that ideally nothing more, besides a confirmation, need be done. In the case of memorization of appointments, the transfer is "quick" because it is *postponed* until a more opportune time.

Section 3 develops those two techniques ("listening in" and "postponing transfer") as the keys to making the Calendar Guardian Agent more usable than current scheduling systems.

3 Method

3.1 Calendar guardian agent

The two techniques for reducing scheduling overhead described in the previous section (*listening in* and *postponing transfer*) can be used to circumscribe a new agent for scheduling appointments. This agent, the CGA, will listen to conversations and attempt to assist the user through scheduling the appointment like a human assistant would. If the user is busy, the CGA can be ignored and it will patiently forward the scheduling to a more convenient time as if the user had committed it to memory.

We are currently building a prototype version of the CGA. The initial system has been built around a SaintSong e-PC Pentium III 700MHz brick-computer running UNIX. The system is carried in a vest by the first author through out the day. A MicroOptical heads-up display is used to place a screen in front of the user so that alerts and feedback during scheduling can be received immediately. We found that a noise-cancelling boom microphone, although bulky, is necessary to get good results when using the IBM ViaVoice speech recognition libraries. One important point is that the touchpad for mouse control has been placed very close to the resting position of the user's hand. This means the mouse can be used without any set-up time while sitting, standing, or even walking.

3.2 Dialog tabs—listening and postponing

The graphical portion of the user interface we are building displays the words the user has recently said for visual feedback. The appointment detection module which performs the *listening in* function is not yet

complete on our prototype. But once it is, as the user speaks appointment times during normal conversation, a form of dialog box, which we are calling a "dialog tab", will pop up. It is non-modal and appears as a small (two pixel wide) tab on the right side of the heads-up display.

If the user is not paying attention to the screen at the moment, these dialog tabs are too small to be distracting. However, that doesn't mean they are difficult to use. We have taken 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 one it should be easy to do so even in poor motor control situations such as while walking and talking. Additionally, the tabs are stacked in order of arrival and more recent tabs are longer; this make the last uttered appointment especially easy to hit.

Dialogs may be dismissed without even opening them first by right clicking. Hovering over a tab shows the time span recognized. A left click will immediately open the calendar to the mentioned date and time allowing the user to fill-in or correct any fields before commiting the appointment to the reminder system. If the user ignores the tabs they remain on the side of the screen until the user has the time to make scheduling her primary task. In this way, the user can *postpone transfer* of the knowledge in her head that is needed to finalize the appointment. However, unlike plain memorization, the user will be less likely to forget about scheduling appointments deferred in this way because the dialog tabs will jog her memory when she takes a free moment to glance down.

4 Challenges

There are three major issues that must be addressed. Voice recognition is not perfect, other's privacy must be respected, and active agents, when poorly implemented, can increase cognitive load.

4.1 Voice recognition is unreliable

In our experience, current voice recognition techniques, while adequate for dictation in an office environment, suffer severe degredation in mobile environments. For example, our system loses accuracy when given the rapid, clipped speech used during conversations. It also suffers from a high number of spurious insertion errors ("false alarms") from ambient noises (e.g., wind) and can become unusable even with a noise-cancelling microphone. We do not believe there is a single solution to the problem of unreliable voice recognition in an unconstrained environment. Instead we will investigate using several techniques in concert.

If the user is aware that she has just or is just about to orally schedule an appointment she can assist the recognition in the following three ways. First, the user can manually signal the system to "pay closer attention" (that is, to lower the normally high threshold of confidence required to detect a scheduling event). Second, the user has constant visual feedback available of what the voice recognition system thinks she is saying and can repeat misrecognized phrases. Third, the user can adapt her speaking patterns to the grammar she knows has worked with the CGA in the past.

To mitigate the effects of imperfect voice recognition, we are designing the CGA's user interface to have a low-cost when mistakes occur. Not only should the display be non-distracting but it should be easy for the user to dismiss spurious errors or correct entries that are partially right. (See 4.3 for more on this). One of the standard techniques to lower the probability of voice recognition errors is to use language model subsetting [Schmandt, 1994] to restrict the application's domain with a grammar of appointments. Speech recognition software often allows a programmer to examine the "confidence levels" for the individual words recognized. Another standard technique is to set a high threshold on the recognition system to throw out questionable words. A possible path of research is to use a limited version of "topic spotting" so that the recognition system can detect conversations in which it is likely that an appointment will be scheduled, and lower the threshold accordingly.

4.2 Privacy

While some researchers have suggested that wearables can be used to protect a user's privacy [Feiner, 1999], a wearable with recording devices may invade other's privacy [Strubb et al., 1998]. It is important to address these privacy concerns in the design of the system. Even if privacy were not an issue in itself, audio recording

of conversations without consent is a legally murky issue. In fact, fourteen states in the United States have laws requiring the consent of all parties in certain situations. Since our system is meant for everyday life, it is almost certain to be used in situations where conversers should have a reasonable expectation of privacy.

The primary method we use to preserve privacy is a "noise-cancelling" microphone which attenuates speech from other people to an essentially inaudible level. We are also currently only saving the text transcript from the voice recognition system; this removes even the possibility of later amplifying the audio recording to discern those near inaudible murmurings. However, if the CGA only has one side of the conversation, some scheduling events can be much more difficult to detect correctly (or at all). We are using two methods to mitigate this. First, the user can assist the CGA by repeating key scheduling times and important points that another speaker has put forth. Results of preliminary tests of this technique are encouraging: since repeating what another person said as a form of confirmation is a standard social custom, few people notice that the user is repeating for the benefit of the wearable (if they notice at all). If we assume the other party also has a networked wearable computer performing speech recognition, the two wearables can negotiate to swap transcripts. The model of swapping transcripts instead of recording other people directly may also improve voice recognition accuracy because each user will have a fixed microphone position and a speech recognition system trained to their voice.

4.3 Cognitive load and Attention

Cognitive load refers to the total amount of mental activity imposed on working (short-term) memory at an instance in time [Sweller, 1994]. The major factor that contributes to cognitive load is the number of elements that need to be attended to. Miller [Miller, 1956], gives the threshold of working memory to be 7 ± 2 items. Because working memory is so limited, if the CGA is to be of any use it must present an extremely low cognitive-load interface to the user. That is, the CGA must keep out of the user's working memory (as far as possible) even though it is potentially visible during all waking hours of the user's life.

To reduce the cognitive burden our system is designed to be easy to ignore and dismiss. The system will show new information only rarely. This is possible because the user is able to assist the voice recognition (see section 4.1), and thus the default threshold can be rather conservative. Additionally the system uses a "ramping interface" [Rhodes, 2000] in the form of dialog tabs (3.2) to show progressively more information to the user yet always provide an easy way out (by a right-click). Using a ramping interface also has the benefit of making errors in voice recognition less costly. Since the tabs are so easily dismissed, the user should be able to manage a higher number of false-alarms from the appointment detection module.

5 Related Work

LookOut [Horvitz, 1999] is an agent which parses the text in the body and subject of an email message, identifies a date and time associated with an event and attempts to fill in relevant fields in an appointment book. The system displays its guesses to the user and allows the user to edit its guesses and to save the final result. The LookOut system gives valuable insight into designing an agent interface that can manage uncertainties about the user's goals by collaborating with the user. However, because e-mail messages are already postponable, LookOut does not offer the same benefits as the CGA which works in the domain of oral conversations to help the transfer of knowledge during high cognitive load situations.

Verbmobil [Wahlster, 2000], created by a large consortium from academia and industry, is a speech-to-speech translation system (between German, Japanese, and English) for spontaneous dialogs in mobile situations. Verbmobil can operate in different domains with appointment negotiation being the most relevant to the CGA. Although Verbmobil does not perform any of the scheduling functions of the CGA, it does demonstrate the feasibility of processing spontaneous speech to retrieve appointment scheduling information [Kipp et al., 1999].

CybreMinder [Dey and Abowd, 2000] allows users to create a reminder message for themselves (or someone else) to be delivered using a mobile device when an associated situation has been satisfied. A similar project, the Memory Glasses [DeVaul et al., 2000], is an attempt to build a wearable, proactive, context-aware memory aid based on wearable sensors. Both of these projects are good complements to the CGA; they can deliver the appointment reminders which the CGA generates.

6 Conclusion and Future Directions

	Technique	Cost
PDA	—	Time
Human	Listening in	Money
Memory	Postponing	Reliability
CGA	Both	Some of each

We have shown an outline for a wearable computer agent that can be built to assist with scheduling appointments. Scheduling is a task which could be significantly faster but usually the tradeoffs in unreliability (for memorization) or monetary expense (for a human assistant) are too high. However, we can apply the techniques of "listening in" and "postponing transfer" to create an agent that could strike a balance between the costs.

This proposed system has three major hurdles to clear: unreliable voice recognition, privacy concerns, and added cognitive load. We have described multiple possible maneuvers for each hurdle. Future work includes completing the appointment detection module and evaluating the system as a whole.

References

[DeVaul et al., 2000] DeVaul, R. W., Clarkson, B., and Pentland, A. (2000). The Memory Glasses: Towards a wearable context aware, situation-appropriate reminder system. In CHI 2000 Workshop on Situated Interaction in Ubiquitous Computing.

[Dey and Abowd, 2000] Dey, A. K. and Abowd, G. D. (2000). Cybreminder: A context-aware system for supporting reminders. In *Proceedings of Second International Symposium on Handheld and Ubiquitous Computing, HUC 2000*, pages 172–186. Springer-Verlag.

[Feiner, 1999] Feiner, S. K. (1999). The importance of being mobile: some social consequences of wearable augmented reality systems. In *(IWAR '99) Proceedings. 2nd IEEE and ACM International Workshop on Augmented Reality*, pages 145–148. IEEE.

[Horvitz, 1999] Horvitz, E. (1999). Principles of mixed-initiative user interfaces. In *Proceedings of CHI'99*, pages 159–166. ACM, Addison-Wesley.

[Kipp et al., 1999] Kipp, M., Alexandersson, J., and Reithinger, N. (1999). Understanding spontaneous negotiation dialogue. In *Proceedings of the IJCAI-99 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 57–64. International Joint Conference on Artificial Intelligence.

[Miller, 1956] Miller, G. A. (1956). The magical number seven plus or minus two : Some limits on our capacity for processing information. *Psychological Review*, 63:81–97.

[Norman, 1988] Norman, D. (1988). *The Psychology of Everyday Things*. Harpercollins.

[Rhodes, 2000] Rhodes, B. J. (2000). *Just-In-Time Information Retrieval*. PhD thesis, MIT Media Laboratory, Cambridge, MA.

[Schmandt, 1994] Schmandt, C. (1994). *Voice Communication with Computers: Conversational Systems*. Van Nostrand Reinhold.

[Shneiderman, 1998] Shneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley, Reading, MA, 3rd ed edition.

[Strubb et al., 1998] Strubb, H., Johnson, K., Allen, A., Bellotti, V., and Starner, T. (1998). Privacy, wearable computers, and recording technology. Panel discussion, The Second International Symposium on Wearable Computers, October 19–20, 1998, Pittsburgh, PA.

[Sweller, 1994] Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction*, 4:295–312.

[Wahlster, 2000] Wahlster, W., editor (2000). *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer-Verlag. Verbmobil is a speaker-independent and bidirectional speech-to-speech translation system for spontaneous dialogs in mobile situations.

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