Augmenting Conversations Using Dual–Purpose Speech

Kent Lyons, Christopher Skeels, Thad Starner
Cornelis M. Snoeck, Benjamin A. Wong, Daniel Ashbrook
College of Computing and GVU Center
Georgia Institute of Technology
Atlanta, GA 30332-0280 USA
{kent, cskeels, thad, cmsnoeck, bbb, anjiro}@cc.gatech.edu

ABSTRACT
In this paper, we explore the concept of dual–purpose speech: speech that is socially appropriate in the context of a human–to–human conversation which also provides meaningful input to a computer. We motivate the use of dual–purpose speech and explore issues of privacy and technological challenges related to mobile speech recognition. We present three applications that utilize dual–purpose speech to assist a user in conversational tasks: the Calendar Navigator Agent, DialogTabs, and Speech Courier. The Calendar Navigator Agent navigates a user's calendar based on socially appropriate speech used while scheduling appointments. DialogTabs allows a user to postpone cognitive processing of conversational material by proving short–term capture of transient information. Finally, Speech Courier allows asynchronous delivery of relevant conversational information to a third party.

Categories and Subject Descriptors: H.5.2 [**User Interfaces**]: Voice I/O, Natural Language, Input devices and strategies

Additional Keywords and Phrases: Speech user interfaces, dual–purpose speech, mobile computing

1 INTRODUCTION

Much of our lives is spent communicating with others: a study of office workers found that 60–85% of their time at work was spent in interpersonal communication [17]. Increasingly, our interactions are in mobile settings; for two office workers, Whittaker et al. found that 17% of their total work day was spent in conversations while "roaming" or away from the desk [27]. In this paper, we present a technique designed to leverage a user's conversational speech. Specifically, we are interested in supporting conversational tasks by utilizing dual–purpose speech. A dual–purpose speech interaction is one where the speech serves two roles. First, it is socially appropriate and meaningful in the context of a human–to–human conversation. Second, the speech provides useful input to a computer. A dual–purpose speech application can listen to one side of a conversation to provide beneficial services.

UIST '04, October 24–27, 2004, Santa Fe, New Mexico, USA.

Many office workers have adopted the practice of carrying personal digital assistants (PDAs) and other mobile computing technology to assist them during conversations. The computers are used to schedule appointments, take notes, and jot down reminders. Manually entering the information encountered during the conversation is the predominant input mechanism. The following example illustrates the issue. Alice is trying to schedule a meeting with her manager, Bob. *Italics* denote the process of Bob using his PDA:

Alice:	Bob, can we meet next week?
	Bob pulls out his PDA.
Bob:	Next week you said?
	Bob starts the scheduling application.
Alice:	Yes, how about Monday?
	Bob uses his stylus to switch to month view.
Bob:	Monday, let me check.
	He selects next Monday to change to day view.
	I'm busy all day Monday.
	Bob advances the calendar one day.
	How about Tuesday?
Alice:	Tuesday at one then?
	Bob selects the 1:00 entry.
Bob:	Sounds good. I'll pencil you in at one.
	Bob enters Alice's name at 1:00 and puts away his PDA.

This example illustrates some of the current interaction issues with tools used during conversation. Although the information to schedule the appointment was spoken during the conversation, Bob must still manually enter it into his computer. Additionally, it is difficult for Bob to participate in the conversation and navigate through the applications on his PDA at the same time. Instead, he must put Alice "on hold" while he interacts with his scheduler.

Our proposed technique of utilizing dual–purpose speech allows Bob to converse with Alice while also providing sufficient information for his computer to automatically move through his calendar. Although our method uses speech recognition, our approach does not require Bob to suspend his conversation with Alice. Instead, our application allows Bob's speech to fulfill its traditional conversational role with Alice while also serving as input to his computer. We explore this example more thoroughly in Section 4.1.

In the remainder of this paper we discuss more in–depth the

idea of dual–purpose speech and explore the issues involved in mobile speech recognition from the standpoints of technology and privacy. Next, we present three applications which utilize dual–purpose speech: Calendar Navigator Agent, DialogTabs, and Speech Courier. We then discuss the common dual–purpose speech issues raised by these applications and describe the common speech infrastructure we have utilized. We then present an evaluation of the Calendar Navigator Agent and finally discuss future work.

2 RELATED WORK

The concept of machine speech recognition was popularized by Vannevar Bush in his 1945 Atlantic Monthly article "As We May Think" [4]. A speech interface was hypothesized to be faster and more "natural" than typing or writing, and initial Wizard of Oz experiments by Gould in 1983 on computer assisted dictation supported this hypothesis [9]. Most public development efforts in the last decade have focused on dictation or interactive voice response systems (e.g. "Show me the flights from Dallas to Pittsburgh") [14]. Several researchers have explored the limits of current speech recognition technology and its appropriateness for various interfaces and situations [20, 31, 16, 7, 19]. Shneiderman provides a brief overview of the issues in his "Limits of Speech Recognition" [22], and Cohen and Oviatt provide a detailed list of conditions when speech may be advantageous in "The Role of Voice Input for Human-Machine Communication" [6].

In this paper, we employ many speech interface techniques described by these authors to constrain our problem of recognizing speech. Our work is also influenced by systems which forgo speech recognition and store the audio directly, using other cues such as pen strokes, location, or time of day for indexing the audio [26, 25, 29, 30, 10]. Such "speech–as–data" systems are directed at situations when the amount of spoken information is overwhelming, such as attending a conference. By using similar interface techniques, our applications degrade gracefully despite potential errors with our speech recognition.

Two applications of interest are Lumiere and the Remembrance Agent. Lumiere models the user and her context to enable applications to provide assistance to the user [12]. Similarly, one use of dual–purpose speech is with applications that act upon the content of a conversation to provide services to the user. The Remembrance Agent (RA) performs continuous associative searches based upon the user's current document and context [18]. In preliminary investigations of dual–purpose speech, we explored using the RA with speech recognition as the data source. While the user could produce dual–purpose speech, the conjunction proved to be inappropriate as the error rates were too high to produce meaningful RA results. This initial investigation prompted us to explore a more constrained domain where we could control the vocabulary to increase speech recognition results.

Work on human–human communication is also relevant to our use of dual–purpose speech. In particular, Speech Acts Theory states that the act of saying something performs an action [2, 21]. In a dual–purpose setting, one utterance might perform two speech acts: one for the conversational partner and one for the computer. In general it would be difficult to automatically interpret speech acts with a computer because the computer has limited access to the user's history and context, and this information is critical to the meaning of a speech act. Furthermore, people often mean more than what they actually say [21]. As we will show, the scope of our applications is sufficiently constrained so that we can make some assumptions about the nature of the speech being used, and all of our applications use push–to–talk so that the user segments the machine relevant portions of the speech.

3 DUAL–PURPOSE SPEECH

Dual–purpose speech may already be familiar to the reader from other settings. For example, a lawyer may have her assistant, Alice, in the office while on the telephone with a colleague. Upon agreeing to exchange some information, she might tell her colleague "My assistant Alice will send you our new proposal today." This utterance is dual purpose; it informs the colleague of the lawyer's intention and provides Alice with the specifics needed to fulfil her instructions without further interaction. We explore this scenario with our Speech Courier application (Section 4.3).

We are extending the concept of dual–purpose speech to be a computer interaction technique. Consider a problem described in 1998 by the Boston Voice Users Group [8]. One of the group members, who used a commercial speech recognition package for his everyday work, noticed that it was inconvenient and socially awkward to disengage the system when guests visited his office. Before he could speak to his guest, he had to turn off the system by saying "Go to sleep." He would then turn to his visitor, say "Just a second" and remove his headset and earpiece. Eventually this individual discovered the solution to his problem. Rather than telling the system "Go to sleep," he changed the stop command for the system to "Just a second." This modification allowed his speech to serve a dual purpose: it disabled the speech recognition system and gracefully informed his guest that he would be ready to converse shortly. The dual–purpose speech transformed a socially awkward situation into one in which a single utterance served two purposes: a command to the computer and a polite comment to the guest.

We have developed this technique as a way to enable computer support during conversations. Effective use of speech as an interaction technique in this domain is challenging. During a human–to–human conversation it is important that any speech interaction with a computer fit the flow of the conversation. For instance, there are numerous situations where it would be socially inappropriate to talk directly to a computer. By using dual–purpose speech, a person can maintain socially appropriate speech: speech where the language and grammar used fits the conversation. While it is important that the language used is socially appropriate it might not be strictly "natural." The user may need to slightly modify her language to effectively use the application. Even so, with our dual–purpose speech the resulting conversation still follows social conventions and sounds "natural" to the conversational partner. The applications we present in Section 4 utilize the content from the user's side of the conversation and attempt to minimize disruptions in the flow of conversation.

One notable feature of our applications is that they depend

only on the speech of one person, the user. Many other projects involving scheduling recognition tasks assume that all sides of the conversation are available [24, 3]. Recording other people's speech without their permission, however, leads to privacy concerns. Also at issue are the limitations of current speech recognition technology.

With only one side of the conversation available, one might think that it is infeasible to obtain all of the required information to complete a task such as scheduling. However, the user can assist the computer by repeating important points that the other person has stated. People often repeat what another person has said to confirm understanding. It is likely that the user already repeats much of the critical information, and the conversational partner is unlikely to realize that the user is repeating any additional information for the benefit of his applications. The example conversation in Section 4.1 reflects this behavior.

3.1 Privacy

A primary concern with speech recognition is the need to record audio, which can lead to issues with privacy. In most areas of the United States, recording of conversations with electronic devices is permissible if at least one participant in the conversation is aware of the recording. In twelve states, however, all participants in a conversation must give consent for recording in most situations.

We have constrained the use of speech in our applications in an effort to preserve privacy. Currently many mobile devices, such as mobile phones, have the ability to record the audio of conversations around the device. Anecdotally, our colleagues have found that it is possible to record people's voices from across a room on some mobile phones. Our primary mechanism for avoiding this effect and insuring the privacy of others is to use a high quality noise cancelling microphone. Worn near the user's mouth, these microphones cancel out nearly all ambient sounds except for the user's voice. In our experience, this greatly reduces the volume of or eliminates the conversational partner's voice from the captured audio. With this technique, our applications utilize only the user's side of the conversation.

3.2 Speech Recognition

Limitations of current speech recognition technology make recognizing meaningful portions of casual conversation very difficult. 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. For instance, many speech recognizers have not sufficiently addressed the varying noise situations that occur during mobile speech. Bursty street traffic noise and microphone noise due to wind can significantly impact a recognition system through insertion errors.

While recognition systems will continue to improve, some errors must be expected. A key strategy we employ to reduce the number of errors is push–to–talk. With push–to–talk, the user specifies which parts of the conversation the computer should attend to by pressing a button. This greatly simplifies the speech recognition task. Instead of continuously processing speech, the computer only needs to recognize the portions of a conversation marked by the user. These phrases contain higher ratios of known keywords and sentences to out–of–vocabulary words and out–of–grammar sequences. We can further reduce errors by formulating appropriate grammars and vocabularies to be recognized. Phrases are chosen to cue the applications while simultaneously informing the user's conversational partner in a socially acceptable manner. While these restrictions are not ideal, they enable us to explore the uses of dual–purpose speech and might be eased as technology improves.

4 APPLICATIONS

In this section we present three prototype applications that illustrate our technique of dual–purpose speech. Since many conversations occur while roaming [27], we built our applications so that they can be used while mobile. These dual–purpose speech applications reduce the amount of manual input and instead reuse material from the conversation.

The three applications are the Calendar Navigator Agent, DialogTabs, and Speech Courier. The Calendar Navigator Agent aids in scheduling. DialogTabs enables a user to augment her short term memory. The third prototype, Speech Courier, enables a user to alert a non–present third party to relevant material from her conversation.

4.1 The Calendar Navigator Agent

The Calendar Navigator Agent (CNA) is a calendar application that has been augmented to utilize the user's speech during a social interaction. The CNA automatically navigates a person's calendar based on a socially appropriate dialog used while creating an appointment. The goal is to allow user interaction with the calendar that has minimal disruption of the scheduling conversation.

When the Calendar Navigator Agent is started, it shows a familiar style of scheduling application (Figure 1a). The graphical interface is similar to common scheduling applications available on PDAs or desktops. As the user proceeds with a conversation, he can hold the "talk" button to run the speech recognition. The speech fragment is processed by the speech recognition engine using a limited grammar tailored to calendaring (for more details, see Section 6.1). Specific keywords such as "next week" or "Monday" are recognized by the CNA's speech recognition engine and used to perform specific actions. If an error is made and an improper action is performed, the user can press a single button to undo the last command.

In Section 1, we described a motivating scenario for our work in which Alice is trying to schedule an appointment with Bob. We will now revisit that scenario, and show how using dual–purpose speech eases the conversation for both participants. Bold face text indicates words spoken by Bob while push–to–talk is active.

The conversation begins with Alice requesting a meeting with Bob.

Alice: Bob, can we meet next week?

Bob starts the CNA (Figure 1a) and presses the "talk" key to activate recording.

Bob: **Next week you said?**

Bob releases the "talk" key to stop recording. The CNA recognizes the key words "next week" in the sentence; knowing the current date, it jumps the display to next week (Figure 1b). As this is occurring, Alice is speaking:

Alice: Yes, how about Monday?

Glancing at Monday on the display, Bob quickly sees several meetings, but it's unclear for how much of the day he'll be occupied.

Bob: **Monday?** Let me check.

The CNA recognizes the keyword "Monday," and switches the view to a close–up of Monday (Figure 1c). It is now clear to Bob that Monday is mostly full. Remembering from the week view that Tuesday seemed clear Bob suggests to Alice:

Bob: I'm busy all day Monday.
How about Tuesday?

The CNA, detecting the keyword "Tuesday," jumps the view to the next day (Figure 1d). Bob can see that he has few appointments. Alice suggests a time.

Alice: Tuesday at one, then?

Bob sees that one o'clock on Tuesday afternoon is free.

Bob: Sounds good. **I'll pencil you in at one.**

The CNA recognizes "one" as a time and creates a new appointment (Figure 1e). Bob may now finish the conversation with Alice. Afterwards, he can fill in the rest of the relevant information for his meeting at his leisure as our speech recognition engine is currently not capable of recognizing the names of people or places.

This conversation is nearly the same as the original; however in this scenario, the amount of information that Bob has to manually enter into the schedule is greatly decreased. Instead, the CNA uses conversational information to navigate the calendar. Bob's interaction with his computer is reduced to using the push–to–talk button, pausing briefly during the conversation to glance at his calendar, and filling in the uncaptured meeting information after the conversation is over.

4.2 DialogTabs

In the previous example, we show how the CNA allows navigation through a calendar. Bob postponed the job of filling out the details in his scheduler entry until after his conversation was over. A natural extension of the CNA would be to capture the audio for this portion of the conversation. The idea of postponement during a conversation is explicitly supported with our next application, DialogTabs.

DialogTabs is designed to help compensate for the limits of short–term memory. Unlike other short term audio reminders (such as the Personal Audio Loop [10]) DialogTabs only processes the user's side of the conversation and uses a push–to–talk button to segment out the relevant portion of a conversation. A small widget, the Dialog Tab, is created to provide a visual reminder of the recording. After the conversation, the user can re–listen to the postponed audio and view an attempted speech–to–text translation (Figure 1f).

Imagine that Bob, after finishing setting up his meeting with Alice, encounters his boss Eve in the hall as she is on her way to an important meeting. Eve has some information for Bob: she wants him to call a client and quickly tells him the phone number.

Eve: Bob, please call our client about the new proposal. They are out of the office; the number is 555-1292.

Rather than open the notepad application on his PDA and try to write the number or look for a pen and a scrap of paper, Bob quickly pushes the DialogTabs button and repeats the number back to Eve.

Bob: **555-1292**. I'll do it now.

When Bob stops recording, DialogTabs creates an unobtrusive tab on the side of Bob's screen; as Bob returns to his phone he can go back and view the tab with the number to make the call. In addition to recording the phone number, Bob exhibits good social etiquette; by repeating the number back to Eve, Bob lets Eve know he heard her correctly.

DialogTabs is explicitly designed to make use of dual–purpose speech. While it could be used as a general short term audio reminder outside of a conversation, using dual–purpose speech makes it well suited as a conversational aid. Many conversations are very short and any time spent diverting attention towards a PDA or paper takes away from the conversation. By reducing the interaction to a single button press and reusing speech from the conversation the cost of the interaction becomes very low.

Visual feedback for each speech segment is generated by showing a Dialog Tab. As they are created, tabs stack vertically in order of arrival. The most recently created tab is the tallest, appearing at the top of the stack and covering twice as much screen space as the next tab. Together the tabs appear as a thin vertical bar at the right edge of the display (Figure 1f). During the course of the day, several tabs may queue up, but the user does not need to process them until he has the time and inclination to do so. The stacked tabs provide a reminder of the information that is waiting for attention, so the user can postpone considering the conversational segments without fear of forgetting them. As each tab is created, the system attempts to recognize the segments of speech recorded for each tab. Hovering the mouse over a tab displays the recognized text, while clicking a tab brings up a dialog box showing a visual representation of the recorded audio along with the text (Figure 1f). The user can click on portions of the audio or words in the text to hear that segment of audio.

Creating a grammar for a general purpose DialogTabs application would be very challenging. To address this issue, we have built several different versions of DialogTabs that use task–specific grammars. Our first uses the CNA grammar while another uses a grammar designed to parse phone numbers. However, even in a more general unconstrained case, DialogTabs is designed to be useful with numerous

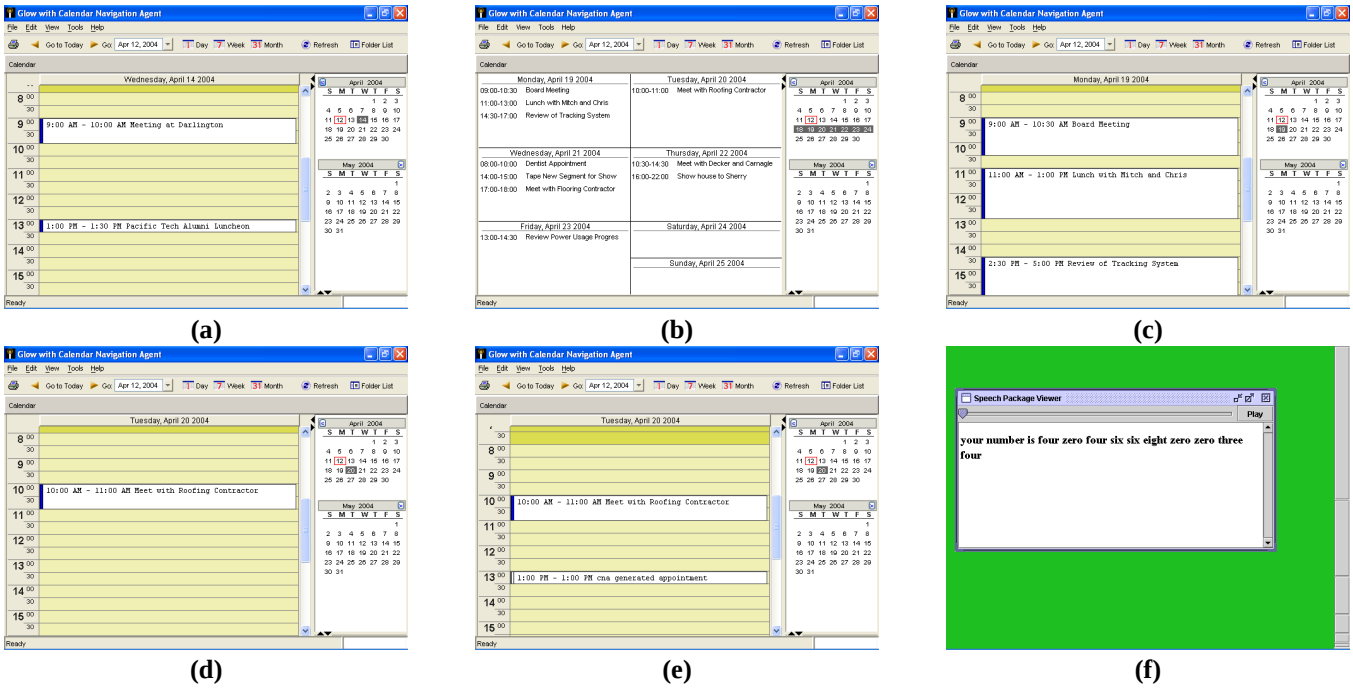


Figure 1: (a) The CNA starts and displays the current date. (b) Cued by "next week," the CNA shows the overview of Bob's schedule the following week. (c) The CNA recognizes "Monday" and shows the detail view for that day. (d) The CNA jumps forward one day when "Tuesday" is recognized. (e) Once the CNA recognizes the time, one o'clock, a new appointment is created. (f) DialogTabs display unobtrusively on the right side of the display. The pop–up allows the user to see the transcribed speech and listen to portions of the audio.

recognition errors. An inaccurate transcript can be sufficient to remind the user of the contents of the conversation fragment, and if not, the user can replay the original audio. Our graphical interface for the transcript is similar to that of the SCANMail system [28], which allows users to visually browse voicemail messages.

4.3 Speech Courier

Our final prototype application is Speech Courier. This tool is designed to relay relevant conversational information to an absent third party and was inspired by informal observations of a high level manager and his work routine. Communication and delegation of tasks to the manager's coworkers consumes much of his work day. Several times a day while conversing with a colleague, either face–to–face or on the telephone, a new task for his assistant is generated. Often his assistant is present during the conversation waiting for tasks that might be created.

For example, Eve might say to Bob:

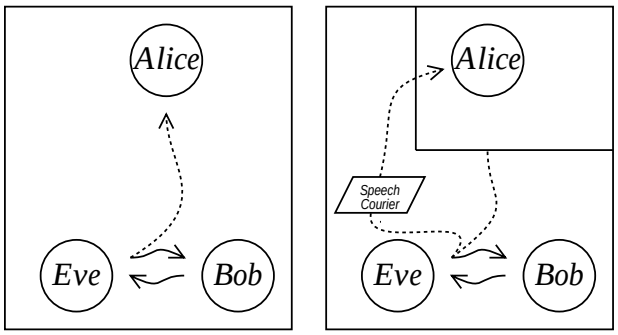
Eve: Yes Bob, Alice will email you the write–up for our new proposal.
Bob understands he will get an email.
Alice knows to send the email.

Alice is present during the conversation and Eve's speech serves a dual purpose: informing Bob and tasking Alice. Figure 2a depicts this situation. Alice understands the new task even though there was not a separate explicit communication between Eve and Alice. Unfortunately this type of interaction requires Alice to be present for the conversation and limits her ability to do other work. If Alice is not present, Eve needs to remember at some other point to give her the new task. As Eve is very busy and often gets distracted by other work, she can easily forget to assign the task to Alice. With the manager we observed this happen on several occasions.

Speech Courier can be used to transform the synchronous dual–purpose face–to–face speech of this situation to a remote asynchronous communication. Using Speech Courier, a user can easily capture an important part of the conversation and send it to a non–present third party. The user marks the important points of the conversation using the push–to–talk button as with our other applications. Once the audio is captured, the speech recognition engine generates a transcript and the audio and transcript are bundled into a package and sent to the third party recipient via email. In our implementation, a single "assistant" user is configured to receive the package and they might be used to convey action items, tasks, reminders, or updates to the non–present person.

Returning to our example, if Alice were not present she would not overhear her task. Using Speech Courier, Eve can tag and save the relevant portion of her conversation and send it to Alice. During the conversation with Bob, she uses the Speech Courier button to record the relevant portion of her speech.

Eve: **Yes, Bob, Alice will email you the write–up for our new proposal.**
Bob understands he will get an email.
Speech Courier sends the task to Alice.



(a) **(b)**

Figure 2: (a) When present, Alice can follow the conversation between Eve and Bob waiting for tasks. (b) When Alice is absent, Eve saves relevant portions of her conversation with Bob using Speech Courier, which then forwards the information to Alice.

Speech courier creates a package with the audio and a transcript and automatically sends it to Alice (Figure 2b).

Our speech recognition language model for Speech Courier is more broad than CNA or DialogTabs, but still rather limited. The speech recognition produces several errors when words are not in the vocabulary. The recorded speech would likely be sufficient to understand the message but the addition of a mostly correct transcript should improve the utility of the application [28]. Because this information is intended for another person, the user might wish to correct any errors or add additional comments. Speech Courier provides the user the ability to edit the recognized text before sending the package and uses an interface similar to the pop–up of DialogTabs (Figure 1f). As the package is created when the user is engaged in conversation, she would likely delay this interaction until finished. The editing capability allows the transcript to serve as a rough draft for a note destined to the third party.

5 DISCUSSION

Calendar Navigator Agent, DialogTabs and Speech Courier all use dual–purpose speech to provide support for a user engaged in conversation. This technique eliminates, postpones, or reduces manual interactions until socially appropriate. While ease and speed of interaction are design considerations in any application, the duration of an interaction is critical when designing tools for use in conversation. Conversations in the office are quite brief; one study shows the average length of a conversation as a mere 113s, while 50% of conversations last less than 38s [27]. Any time spent interacting with a computer can disrupt the flow of conversation. In the worst case, the user will avoid using the tools altogether.

Our dual–purpose speech applications have the advantage of a low cost of failure, where failure is many recognition errors on the part of the speech recognition engine. With DialogTabs and Speech Courier, a completely inaccurate transcript will mean only that the user must listen to the entire clip of speech which can be done after the conversation is completed. Imperfect recognition on the part of the Calendar Navigator Agent forces the user to address the error. She must either repeat the phrase in a socially appropriate way while avoiding a cascade of errors or revert to manually navigating through the calendar. In the latter case, the rest of the interaction would be the same as if she had used manual input the whole time.

	Restricted	Unrestricted
Intentional	CNA, DT, SC	Co–located assistant
Unintentional	N/A	Naive ideal

Table 1: Design matrix of dual–purpose speech. Our applications are restricted and intentional.

Even though our applications only listen to the user's side of the conversation to protect the privacy of others, they still offer beneficial functionality. In the cases where the dual–purpose speech applications need additional information, the user can repeat back to her conversational partner. The echoing of key dates in the CNA or repeating a phone number during a conversation allows the user to both give input to the computer and also confirm that the message has been heard and understood properly. Repeating key points is often already performed when communication channels are poor or the information is particularly important. For instance, military radio conversations have special protocols to ensure proper communication [1]. Even though repeating back information for use by the computer may mean a small change in communication habits, the privacy benefits of only recording the user's side of the conversation are significant.

5.1 Dual–Purpose Speech Design Space

These three applications highlight important aspects of dual–purpose speech. The first issue is if the dual–purpose speech is intentional or unintentional. That is, does the user intentionally speak to both her conversational partner and the computer or just to the conversational partner? Closely related is the question of the language used; it can either be restricted or unrestricted. The next issue concerns the intended recipient of the speech. The recipient can be a computer or a person, and if it is a computer, it can act upon the speech like the CNA or only passively record and transcribe the speech.

First we will discuss intentional and restricted dual–purpose speech. Intentional dual–purpose speech is when the speaker intends for her speech to be directed towards both parties. Unintentional dual–purpose speech is formulated only for a single recipient even though the second is listening and acting. Unrestricted language is natural everyday speech with no boundaries, whereas restricted language requires a predefined limited vocabulary (Table 1). All three of our applications are intentional and use restricted language. At the very least, the user must press the push–to–talk button to segment her speech. She must also intentionally restrict her speech to the language model of the three applications. An example of intentional and unrestricted dual–purpose speech can be found in scenario that inspired Speech Courier where Eve talks to Bob while Alice is listening (Figure 2a). Eve is explicitly talking to Bob but also formulating her speech so Alice understands. The case of unintentional restricted dual–purpose speech cannot exist because a speaker can only restrict her speech intentionally. Lastly, unintentional unre-

stricted speech would be the least burdensome for the user, possibly creating a better user experience. Unfortunately current speech recognition requires some restriction in the user's language to achieve satisfactory results. Furthermore, it is not clear if the user's language would contain enough information to be of use given the implications of Speech Act Theory [2, 21].

The next issue to consider is the intended recipient of the dual–purpose speech information and whether an application acts upon the person's speech or uses it passively. DialogTabs is an example of a passive application that buffers the audio and associated transcript for the user. The recipient of Speech Courier package is a non–present third party. The CNA is an example of an application where the computer is the intended recipient of the speech; the user's speech is mapped directly to actions performed on the calendar.

The intended recipient changes the impact of speech recognition errors. The CNA operates directly on the speech, and an error in speech recognition can result in an improper action with the calendar. If there is an error, the user cannot continue with scheduling until it is addressed. To correct errors, the user must divert attention away from the conversation, thereby disrupting the flow and distracting the user from her primary task. In contrast, if a person is the recipient such as with Speech Courier or DialogTabs, an error "only" results in an improper transcription. Errors made in applications such as these do not interrupt the conversation and can be corrected during a less demanding time (after the conversation is over). The value of the transcript decreases as the associated errors increase, but the applications still function and remain useful.

6 IMPLEMENTATION

The implementation of our dual–purpose speech applications requires a mobile computer such as a wearable or laptop capable of performing speech recognition in near real time and the user must wear a high quality noise canceling microphone. We also use an input device for push–to–talk and a display for visual feedback. In this section, we discuss how we met these requirements in building the Calendar Navigator Agent, DialogTabs and Speech Courier.

Utilizing a high quality audio source helps to improve the accuracy of speech recognition and ensures that recorded speech is intelligible when played back. To this end, we used a VXI Talk Pro Max microphone which features active noise canceling and voice enhancement. The noise canceling feature filters out nearly all ambient sounds except for the user's voice, while the voice itself is enhanced by limiting distortion caused by breath pops and other sounds at non–speech frequencies.

For automatic speech recognition (ASR), we used version 4 of CMU's Sphinx software [13]. Sphinx 4 is a highly modular, extensible ASR research system written in Java that has an architecture which allows for the use of custom language and acoustic models. Our prototypes consists of the Sphinx recognition engine, libraries that abstract audio, speech recognition, and visualization services, and graphical user interfaces to these services. All system components are written in Java 2 Standard Edition and run under GNU/Linux and Windows XP. Glow[1], an open source Java calendaring application, was modified for use in the CNA application (Section 4.1). The applications run on a 1.7GHz Intel Pentium IV Mobile CPU laptop and previous implementations of the CNA and DialogTabs have run on an 800MHz Transmeta–based wearable computer.

6.1 Acoustic and Language Models

A key issue in building applications that utilize speech recognition is the use of acoustic and language models. Acoustic models provide information about the low–level features of speech such as phonemes, while language models provide information about pronunciation and grammar.

In general, acoustic models are separated into speaker dependent and speaker independent models. A speaker dependent model will be more accurate for the particular speaker that provided the acoustic data, while a speaker independent model allows many users to be recognized at the cost of reduced overall performance. Given the high barrier of entry for creating acoustic models, we chose to use the freely available speaker–independent DARPA Resource Management acoustic model[2].

An important part of our research was constructing an appropriate language model to use in dual–speech situations. A language model consists of a pronunciation dictionary and a grammar that specifies how words in that dictionary combine. A grammar can specify that a sequence such as "How about we meet next week" is highly probable while the sequence "How a lot of next meet" is not. When a certain conversational task can be assumed such as appointment scheduling, task specific language can be engineered into the grammar to reduce processing time and to achieve higher recognition accuracy. On the other hand, when no specific task can be assumed, a relaxed grammar must be used which is necessarily less accurate.

In our implementation of DialogTabs, we chose the limited task of saving phone numbers. The corresponding language model represents one extreme along the continuum of grammar constraints. A corpus of eighteen sentences and nineteen words was constructed. The corpus contains variants of the phrase "So your phone number is…" and the digits zero through nine. The probabilistic language model generated from the corpus contains 19 unigrams, 36 bigrams, and 26 trigrams.

Though still a constrained task, a much more general language model was built for the CNA. The corpus was prepared by observing the language used by participants during a previous study on mobile calendaring [23]. Example phrases include "How about the day after?" and "Let's meet October twelfth." The corpus contains a total of 1007 phrases and the resulting probabilistic language model contains 121 unigrams, 461 bigrams, and 744 trigrams.

We observed that despite the variation in the language used in the calendaring corpus, there is little variation in the in-

[1] http://groupware.openoffice.org/glow/
[2] http://www.ldc.upenn.edu/Catalog/

tent of the language. We identified three semantically distinct units that could be leveraged for calendar navigation. These are the initial check of a certain date, the subsequent access of other dates when the initial check fails (e.g. the user has a previous engagement), and the final act of confirming the appointment. After recognition is performed on a sentence, keyword matching is applied to determine which of the three actions is intended. For example, finding "March" and finding "20th" would signal the check of "March 20th," even if the spoken sentence was "let's meet in March...how about the 20th?". This keyword–to–intention mapping helps the Calendar Navigator Agent be more flexible in its recognition especially if the user strays outside the language model. This technique in turn reduces the effect of recognition errors and helps to avoid the cost arising from incorrect navigation.

In contrast to the other applications, a more general purpose grammar was used as the starting point for Speech Courier's language model. This was done to explore the use of unconstrained speech recognition in conversational situations that are hard to formulate. Our assumption was that any language model we could construct would perform poorly in an arbitrary situation not accounted for by the model. Our approach was to build a base model that could be iteratively extended according to personal experience, informal observations and future formal usage studies. The base corpus includes one thousand common words as well as the scheduling scenario corpus identified in the CNA language model for a total of 2042 phrases and 1050 words. The probabilistic language model also contains 2437 bigrams, and 1779 trigrams.

7 EVALUATION

We conducted a preliminary study of the Calendar Navigator Agent to investigate its effectiveness for scheduling appointments and for ease of use. Specifically we are interested in the effectiveness of our speech recognition, the ease of use of push–to–talk, and the users' ability to employ the restricted grammar. We focused on the CNA because speech recognition errors are the most critical in this application of our three and scheduling allows us to explore dual–purpose speech with a straightforward and realistic task for users.

7.1 Procedure

Three people from our laboratory used the CNA for this study. Everyone had a passing knowledge of dual–purpose speech and our applications before the study; however, no one had any experience with our prototypes. The trials lasted 60–90 minutes for each person. The CNA ran on a laptop at a desk, and the laptop's screen displayed the application.

The study consists of four parts: a baseline evaluation of speech recognition, a demonstration of the CNA, training in two phases, and finally a test with scheduling appointments using dual–purpose speech. These steps are designed to gradually introduce the users to the language and abilities of the CNA. After the experiment, we administered a questionnaire and conducted an interview.

First using a testing application, we obtain a baseline of each user's speech recognition rates for the language of the CNA. Each user reads through 20 sentences used by the CNA. For each sentence, the subject uses push–to–talk and speaks the presented phrase. If there is an error, they repeat the phrase until it is correct.

	A	B	C	Mean
Accuracy	79.9%	97.5%	83.6%	87.0%
Correct	88.9%	100%	91.0%	93.3%

Table 2: Word level percent accuracy and percent correct for three users.

The researcher next demonstrates the CNA. He navigates and schedules two appointments using speech. The user is instructed to listen to the speech and watch the resulting actions taken by the CNA.

Next are the two training phases designed to instruct the users on the association between the speech used for the CNA and the actions performed in the calendar. Users schedule two appointments per training phase in a sequence of steps. Each step represents one turn of the user's dialog. Part one of the training is the prompted phase. At each step of this phase, the researcher explains the possible actions that can be taken from the current state in the CNA. The researcher then gives the user a phrase to speak, asks her to repeat it to ensure she understands what to say, and the user speaks the phrase to the CNA. The second training phase is the user generated phase. As in the previous phase, the user schedules two appointments step–by–step. However instead of being prompted with what to say, the user is given a more general goal and asked to generate a phrase to use with the CNA. Once the participant generates a correct phrase, she uses it with the CNA.

The last portion of the experiment is the test phase designed to mimic appointment scheduling conversations. The user is asked to participate in nine scheduling dialogs with the researcher. Using the information in the CNA calendar, the user responds to calendaring requests made by the researcher or initiates a dialog given a high–level goal (e.g. "schedule an appointment next week").

At the conclusion of the experiment, we administered a questionnaire composed of 12 Likert scale questions and used the answers as a basis for a semi-structured interview.

7.2 Results

While limited in scope, the results from our study are promising. Table 2 shows the word–level recognition rates for our three users taken from our initial speech recognition baseline phase of the experiment. Percent accuracy is defined as $\frac{N-D-S-I}{N} \times 100\%$ and percent correct as $\frac{N-D-S}{N} \times 100\%$ where N is the total number of words, D is the number of deletions, S the number of substitutions and I the number of insertions. Overall, the mean accuracy for the group is 87.0%, while the percent correct is 93.3%. While better recognition rates would improve our application, one user performed very well achieving 100% correctness and 97.5% accuracy on our 20 phrases. This result indicates that with an improved or adapted acoustic model, we might be able to enhance our overall recognition rates.

While the word–level speech recognition rates provide an

overall sense of the performance of the application, the actions performed by the CNA are more important. For the testing portion of the study, phrases were successfully recognized and acted upon by the CNA without errors 80.2% of the time. Furthermore, each task in the scheduling dialog was completed with at most one recognition error 97.8% of the time. This result implies that uttering the phrase again seems to be effective for the CNA. Our current language model is very limited and was not designed to enable socially graceful correction of errors through speech. For our experiment, the user was asked to repeat a phrase until the CNA performed the correct action. Given this result, we are exploring ways to modify our language model so that a user can repeat or rephrase what she said. This ability would enable the computer to try again, while at the same time minimizing any disruption in the flow of the conversation. Most of the speech recognition errors in our study resulted in no action taken by the CNA as opposed to the incorrect action. It is possible that using a slightly more intelligent algorithm to interpret the speech might increase the ability of the CNA to perform the correct action when speech recognition errors are made.

The questionnaire and interviews provide additional insight. Our users quickly accommodated to using push–to–talk and rated it as fairly easy to use during the scheduling conversation. Our users thought that the language for scheduling with the CNA was fairly acceptable and socially appropriate. Even with the training given, the users indicated that language generation is the hardest part of using the CNA. This issue was demonstrated most clearly when the user initiates the scheduling dialog and cannot simply respond to the conversational partner. The users also realized their own limitations, and this quote is typical: "This app. would be really useful given more training." Ideally, the use of dual–purpose speech should be much more effortless. Our results imply that the users needed more training to become experts in generating the needed speech during conversational situations. We might also attempt to make applications adaptive so that the user's speech can more closely match her natural language. Even with the current limitations, all three users were enthusiastic about the CNA application and agreed using conversations and dual–purpose speech as a means to schedule appointments would be useful.

8 FUTURE WORK

There are several areas we would like to explore in the future. First, we would like to further examine ways to reduce the number of recognition errors. By creating better acoustic and language models which better mimic the everyday settings our applications are designed for, we believe that higher recognition accuracy can be achieved. One option is to use speaker–dependent speech recognition. While this would require a great deal of time from the user to properly train the recognition engine, the increased accuracy and long term benefits might justify the cost. We are also exploring other speech recognition engines which might have better models.

Codifying the language used in everyday situations into appropriate vocabularies and grammars is an interesting research opportunity. The results must reduce the perplexity of the speech recognition problem while maintaining the socially appropriate patterns of natural dialog. One can imagine an effort similar to the DARPA Airline Travel Information Service (ATIS) task where researchers try to capture the "natural" vocabulary and grammar related to a specific task and then create a system that allows very flexible interaction while still being specifically tuned to the task [11, 14, 15].

As discussed previously, due to privacy concerns we currently only use one side of a conversation. One interesting possibility for overcoming this limitation while preserving privacy is for the participants of a conversation to share their one–sided recordings with the other participants, allowing the reconstruction of the entire conversation. Privacy of the shared conversation fragments would be a great concern and would have to be considered carefully. This goal could be accomplished automatically, by providing some sort of authentication between users' devices, and by either time–stamping recordings or by looking at turn–taking behavior [5].

The results from our initial study are encouraging and we are interested in exploring the usability and usefulness of the dual–purpose speech technique in more detail. Of particular interest is the relative costs of using dual–purpose speech in terms of time and cognitive load compared to more traditional mobile devices and interface techniques. We are interested in examining the benefits and drawbacks of using push–to–talk and our currently limited grammars while controlling for speech recognition errors and exploring the design space of dual–purpose speech more thoroughly.

Finally, we would like to deploy our dual–purpose speech applications for long term use in real–world situations and we are interested in exploring the usefulness of the CNA, DialogTabs and Speech Courier in their current limited forms. We are in the process of porting the software to our wearable computers and are exploring the possibility of using the new generation of high–end PDAs, such as the Sharp Zaurus SL-6000, as platforms for these applications.

9 CONCLUSIONS

We introduced the concept of a dual–purpose speech interaction: socially appropriate speech that provides meaningful input to a computer. We show that dual–purpose speech can be employed by applications to augment conversations. Our three applications, the Calendar Navigator Agent, DialogTabs, and Speech Courier, explored this design space, and we identified three aspects of dual–purpose speech: restricted language, intentional use of speech, and intended recipient. We discussed issues of designing interactions based only on the user's speech to ensure privacy and robustness in the presence of speech recognition errors. With future improvements to speech recognition, we expect dual–purpose speech to become more widely applicable for mobile computing.

ACKNOWLEDGMENTS

This material is supported in part by the NIDRR Wireless RERC and NSF Career Grant #0093291.

REFERENCES

1. Allied Communications Publication. *Communication Instructions Radiotelephone Procedures*, September 2001.

2. J. L. Austin. *How to do Things with Words*. Harvard University Press, 1962.

3. S. Busemann, T. Declerck, A. K. Diagne, L. Dini, J. Klein, and S. Schmeier. Natural language dialogue service for appointment scheduling agents. Technical Report RR-97-02, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, 1997.

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

5. T. Choudhury and A. Pentland. Sensing and modeling human networks using the sociometer. In *Proceedings of ISWC*, October 2003.

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

7. C. Danis, L. Comerford, E. Janke, K. Davies, J. DeVries, and A. Bertrand. Storywriter: A speech oriented editor. In *Proceedings of CHI*, pages 277–278, New York, April 1994. ACM.

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

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

10. G. R. Hayes, S. N. Patel, K. N. Truong, G. Iachello, J. A. Kientz, R. Farmer, and G. D. Abowd. The personal audio loop: Designing a ubiquitous audio-based memory aid. In *Proceedings of Mobile HCI*, 2004.

11. C. T. Hemphill, J. J. Godfrey, and G. R. Doddington. The ATIS spoken language systems pilot corpus. In *Proc. of the Speech and Natural Language Workshop*, pages 96–101, Hidden Valley, PA, 1990.

12. E. Horvitz, J. Breese, D. Heckerman, D. Hovel, and K. Rommelse. The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In *Proceedings of Uncertainty in Artificial Intelligence*, 1998.

13. X. Huang, F. Alleva, H. wuen Hon, M.-Y. H. andKai Fu Lee, and R. Rosenfeld. The Sphinx-II speech recognition system: An overview. *Computer, Speech and Language*, pages 137–148, 1993.

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

15. E. Levin, R. Pieraccini, and W. Eckert. A stochastic model of human-machine interaction for learning dialog strategies. *Trans. on Speech and Audio Processing*, 8(1):11–23, 2000.

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

17. R. Panko. Managerial communication patterns. *Journal of Organisational Computing*, 1992.

18. B. J. Rhodes. *Just-In-Time Information Retrieval*. PhD thesis, MIT Media Laboratory, Cambridge, MA, May 2000.

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

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

21. J. R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.

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

23. T. E. Starner, C. M. Snoeck, B. A. Wong, and R. M. McGuire. Use of mobile appointment scheduling devices. In *Proceedings of CHI*. ACM Press, 2004.

24. M. Stede, S. Haas, and U. Küssner. Tracking and understanding temporal descriptions in dialogue. Verbmobil-Report 232, Technische Universität Berlin, October 1998.

25. L. Stifelman. Augmenting real-world objects. In *Proceedings of CHI*, New York, 1996. ACM.

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

27. S. Whittaker, D. Frohlich, and O. Daly-Jones. Informal workplace communication: what is it like and how might we support it? In *Proceedings of CHI*, pages 131–137. ACM Press, 1994.

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

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

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

31. N. Yankelovich, G. Levow, and M. Marx. Designing SpeechActs: Issues in speech user interfaces. In *Proceedings of CHI*, pages 568–572, New York, 1995. ACM.