RC24113 (W0611-153) November 27, 2006
Computer Science

IBM Research Report

Examining Modality Usage in a Conversational Multimodal Application for Mobile e-Mail Access

Jennifer Lai, Stella Mitchell
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Christopher Pavlovski
IBM Corporation
348 Edward Street
Brisbane, QLD 4000
Australia

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Examining Modality Usage in a Conversational Multimodal Application for Mobile e-Mail Access

Jennifer Lai, Stella Mitchell
IBM T.J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532
jlai, cleo, @us.ibm.com

Christopher Pavlovski
IBM Corporation
348 Edward Street
Brisbane, QLD 4000 Australia
chris_pav@au1.ibm.com

ABSTRACT

As Third Generation (3G) networks emerge they provide not only higher data transmission rates but also the ability to transmit both voice and low latency data within the same session. This paper describes the architecture and implementation of a multimodal application (voice and text) that uses natural language understanding combined with a WAP browser to access email messages on a cell phone. We present results from the use of the system by users as part of a laboratory trial that evaluated usage. The user trial also compared the multimodal system with a text-only system that is representative of current products in the market today. We discuss the observed modality issues and highlight implementation problems and usability concerns that were encountered in the trial. Findings indicate that speech was used the majority of the time by participants for both input and navigation even though most of the participants had little or no prior experience with speech systems (yet did have prior experience with text-only access to applications on their phones). To our knowledge this represents the first implementation and evaluation of its kind using this combination of technologies on an unmodified cell phone. Design implications resulting from the study findings and usability issues encountered are presented to inform the design of future conversational multimodal mobile applications.

Categories and Subject Descriptors

H5.2. Information Interfaces and Presentation: User Interfaces, H4.3. Information System Applications: Communication Applications.

Keywords

Multimodal interfaces, modality usage, natural language understanding, mobile phones, speech technologies

1. INTRODUCTION

The cell phone is a very popular device, growing in numbers and popularity worldwide, with an increasing number of functions being added. Given the ubiquitous presence of cell phones, application designers and providers are well served to understand the interaction issues associated with mobile application usage. Since cell phones have neither a large screen nor a highly usable keyboard, input and output modality issues are an important design consideration when creating an application that will be used on a cell phone.

Interaction with a phone is limited by difficulties of input. While researchers have compared efficiency gains between a predictive text input mechanism such as T9 (Grover et. al, 1998) and multi-tap input mechanisms (e.g. James, 2001), the fact remains that most users only resort to inputting text from the keypad of cell phone under limited circumstances. The first is when a short but critical piece of text needs to be entered into the phone; for example, associating a name with a telephone number. The second is the practice of sending and receiving Short Messaging Service (SMS) messages. SMS messaging represents a large and growing trend that is rapidly becoming an established practice in the United States among certain user populations (e.g. teenagers). In Europe (e.g. Germany and U.K.) and in Asia (e.g. Japan, Philippines) there is a strong reliance on this form of messaging. However, SMS users practice a unique form of abbreviated language, creating almost a language of their own in order to limit the amount of text that needs to be entered through the phone (Longueuil, 2002). Examples in English include: Great GR8, Late L8, Later L8R, Please call me PCM, Talk to you later TTYL, Speak SPK, See you later CU L8R, Thank you THKS, Today 2DAY, Tomorrow 2MORO, Want to WAN2. Many of these abbreviations are considered acceptable language for text messaging but have not become mainstream in business related email messages.

Speech technology has been proposed as a suitable alternate text input mechanism for cell phones; however the performance is often degraded in mobile environments (Salonisdis et. al,, 1998, Oviatt, 2000). For instance, speech recognition error rates tend to increase when used on a cell phone due to the wireless network audio encoding quality (Salonisdis et. al, 1998) and the variable background noise present in the user's environment (Oviatt, 2000). Certain tasks such as retrieving and sending email messages are difficult to accomplish in a speech-only environment due to the volume of messages that most business workers receive combined with the transient and invisible nature of a spoken interface. While good progress has been made in the design of speech-only systems with both land lines (e.g. Yankelovich, 1995) and mobile devices (Sawhney, 2000; Lai, 2002), practice has shown that most business users resort to voice-only access to email only when stranded (e.g. stuck in traffic) or use it for read-only access and do not generate new messages or send replies (Lai, 2002).

Multimodal interfaces, i.e. interfaces that accept at least two input modes, have really only started to be used and seriously researched in the past 20 years (Oviatt, 2003). Multimodal systems present a significant advantage over unimodal systems in that they are accessible by users with a wide range of capabilities and are usable in a variety of environmental settings. They are often viewed as the solution to increasing the robustness and accuracy of speech-only systems (Oviatt, 2003) and appear to be well suited for use in a mobile computing environment given the varying constraints placed on both the user and the recognition technology. They present an array of solutions to the range of problems that exist when using text intensive applications in a mobile environment on a cell phone, and are well suited to the divided-attention state of the user. Multimodal interfaces support the use of redundant information to increase speech recognition accuracy and reduce user frustration since the user can speak replies to messages, or can quietly read longer passages of text.

Creating a highly usable interface for browsing and accessing large amounts of textual information over a mobile phone represents a substantial challenge. It is not unusual for the average knowledge worker to receive over a hundred messages a day. The sequential navigation model that was established for voicemail messages, and that is still applied to many unimodal mobile email systems today, does not work well when applied to large numbers of messages. When viewing the daily onslaught of email messages with a graphical user interface (GUI) users routinely do a visual triage, scanning for messages from people whom are important to them, or for message topics that pique their interest. This triage is difficult to do when relying on auditory output, which is slower than the visual channel. A multimodal interaction model thus appears to be ideally suited to mobile email retrieval because it supports visual browsing, and the combination of modalities can help to improve the robustness of the speech recognition in the very challenging environment of mobile usage.

This paper describes a multimodal implementation for a mobile email retrieval application in a Third Generation (3G) network environment, along with the results of a laboratory trial that was conducted to evaluate the system. The first trial examined the usability of the multimodal system (speech, selection and text) and compared it to an existing unimodal (text-only) system in the same domain.. We also describe the architecture used for the implementation. The email domain was selected because we have prior experience in delivering a speech-only email solution for business workers and because it is a text intensive application and thus a good test case for the value of multimodal interaction on a cell phone. The implementation used an off-the-shelf, unmodified cell phone. To our knowledge, the e-PIM (electronic-Personal Information Management) system is the first fully functioning implementation of a conversational multimodal interface on an unmodified cell phone.

2. PRIOR WORK

Prior research has shown that for spatial domains (e.g. dynamic interactive maps), multimodal input is clearly preferred to unimodal input by users and results in positive performance advantages. Multiple studies by Cohen (e.g. 2000) and Oviatt (e.g. 1996), using both simulation systems and prototype implementations on a desktop and laptop, compared efficiency gains and user preferences with map-based tasks, while varying the input modality. Multimodal input resulted in fewer recognition errors, faster input times and greater user satisfaction when compared to speech-only or GUI-only input. However, this finding can not be disassociated from the difficulty of describing spatial locations with speech only. Oviatt found that speech recognition accuracy was higher when combined with pen input since users' multimodal utterances were briefer, contained fewer complex descriptions and had as few as half the number of disfluencies as speech-only input (Oviatt, 1996). This same study also found that users had a strong tendency to switch modes after a recognition error thus leading to smoother error recovery.

More recently, Gong (2003) compared modality usage for a sales application on a PDA. Fifteen users participated in the study; they were placed in a noisy environment and were given a free choice of using speech or stylus. Speech was used as the primary input modality for entering textual data, while the stylus was used more often for navigation and operating drop-down lists. Again this finding would appear to be closely related to the task since drop-down lists lend themselves more closely to stylus selection than speech.

While most of the prior research on modality issues for multimodal systems has focused on input issues, findings from the field of cognitive psychology relating to human attention capacity can be applied to understanding output issues in multimodal/multimedia systems. Wickens' (1983) research in attentional processing resources has shown that is it easier to process information using two different sense modalities (e.g. visual and auditory) than using two attention channels within the same mode (e.g. thinking and listening). When applied to multimedia output, the findings of Mayer and Moreno (1998) provide strong support for the dual-processing theory by showing that redundant verbal explanations (speech and text) produce greater retention and understanding than either modality alone, since students are able to increase working memory capacity by processing the message in both the visual and the auditory channel. However, it is important to keep in mind when designing mobile multimodal/multimedia applications that these findings do not necessarily hold if working memory is overloaded due to the presence of a simultaneous important second task (Baddeley, 1992) such as driving, or negotiating a crowded sidewalk.

The prior work underscores the fact that the question of modality preference and efficiency gains in a mobile environment with a divided attention state cannot be resolved unilaterally for all tasks and all domains. For most domains, the preferred modes of interaction have yet to be established, and are likely to be dependent on personal preferences and history of success, as well as context of use.

3. E-PIM

The e-PIM project focuses on facilitating mobile communication by providing multimodal access, using natural language speech and graphical browser, to enterprise email and calendar entries from a cell phone. User input can be in the form of text input, GUI selection, or automatic speech recognition. Output is a combination of written text displayed on the cell phone screen and spoken synthetic speech.

3.1 Functionality

Supported functions include reading, sending, forwarding, replying, deletion of email messages, as well as checking calendar entries, and creating appointments. By editing a personal profile, users can set their password and configure certain aspects of the application, such as how many days worth of email messages to retrieve, how to pronounce their name, or how fast the text-to-speech voice should speak.

Users can interact with e-PIM using both voice and the GUI browser. Requests can be made from either modality and the system response is presented on both modalities at the same time. Although all the core capabilities are available in both the audio and visual interface, each modality has certain traits that are only available in that modality. For example, using a spoken command the user can request messages about a specific subject (e.g., “do I have any messages about the seminar”), whereas using the GUI browser the user can only request a listing of all the messages in their inbox. During email creation, the graphical browser provides richer function allowing any recipient name to be entered by text, while the spoken interface only recognizes the set of names consisting of other e-PIM users, personal address-book contacts and senders of messages listed in the inbox. These differences are valuable to the user because they capitalize on the strength of each modality.

3.1.1 Voice Interface

E-PIM's voice interface employs a natural language understanding technology (Jurafsky, 2000) that uses statistical techniques to transform text generated by the speech recognizer into formal language statements that express the meaning of the utterance. This allows the user to be very open with their vocabulary and phrasing. The system also supports mixed-initiative dialog, which allows users to switch to a new task without completing a task that they previously initiated. The following dialog example illustrates this feature.

> **User**: *set up a one hour meeting tomorrow*
>
> **System**: *what time should the meeting start?*
>
> **User**: *do I have any messages from David Smith?*
>
> **System**: *you have three messages from David Smith*

3.1.2 Conversational Speech

Speech-based telephone interfaces available in the commercial market today use varying degrees of directed dialog. Directed dialog, as the name implies, uses a style of system prompts that helps to "direct" the user on what to say next. Users are often presented with spoken menu options from which they can make a selection; thus navigating in a controlled manner until the task is completed. Much of the naturalness and power of speech is undermined when the application relies too heavily on the use of directed dialogs, and the user can feel confined to the passive role of waiting for the system to prompt for a specific answer. A Natural Language interface reduces the user's cognitive load since there are no commands to memorize or hierarchies to navigate. Our domain specific language model for speech recognition includes a vocabulary of about 1000 words, and is trained on approximately 11,000 sentences. Concatenative text-to-speech (TTS) with an English accent (since no Australian accent was available) was used for all of the voice prompts.

Unlike existing commercially available voice systems for email, e-PIM supports random access to any message in the inbox. Messages can be accessed sequentially as well, but this is a rather inefficient way of retrieving email if one is just looking for a specific message. For random access we support several parameters including date, sender, subject keyword, urgency, and ordinal number (e.g., “read me the second message from David Smith”, or “show me all the urgent messages received yesterday”). Calendar entries can also be queried via a number of parameters including, time, time range, date, date range, and type (e.g., “show me my calendar entries from 2 to 4 pm tomorrow”).

When a user queries “Do I have any messages about the seminar?” we translate that to “do I have any messages about X” where X (in this example, X is “the seminar”) is then matched to a dynamic grammar built from the subject lines of all the messages in the user’s inbox. Dynamic grammars, such as this message subject grammar, are constructed and supplied to the speech recognizer at runtime. In e-PIM the message subject grammar (for a particular user at a particular time) is made up of each individual word contained in the subject lines (e.g. ‘meeting’, ‘tomorrow’, ‘morning’) as well as each complete message subject (e.g. ‘meeting tomorrow morning’) . The application makes no determination about the urgency of the message from the text, but relies on the sender of the message to specify the importance level of the message. Thus the query “do I have any urgent messages” checks for messages that have the urgent bit set, and displays the headers for messages that match that criteria. The voice prompt presents a summary of the information (e.g. “you have three new urgent messages”).

In addition to the core functions described above, the interface supports some additional requests that help maintain a productive dialog such as:

- **Guide me**: drops back to a more directed style of dialog to help guide the user through the available choices (e.g. “Would you like to check messages or send a message?”)
- **Help**: presents contextual help
- **Repeat**: repeats the last system response
- **Cancel**: aborts the current operation

3.1.3 Visual Interface

The visual interface for e-PIM provides four primary choices on the main menu screen: Check Messages, Send a Message, Check Calendar and Help. These correspond to the core functions that are also available through the voice interface. The one exception is the calendar entry creation function which was only fully available through the voice interface at the time of the pilot and thus was not part of the user trial.

The user activates a choice by selecting it through use of the four-way scroll button on the phone. For most requests, the visual response is displayed all at once, with scrolling used as necessary to access all the information. However, in order to best display the response to "check messages" on the small screen, the email headers contain only sender and subject (see Figure 1). The body of the message is viewable by selection, with additional details (e.g. date and time of the message) being available by click- through.

For interactions that support text entry such as the "send email" screen, the user can opt to enter text rather than speaking by using the keys on the telephone keypad with the multi-tap method. This method requires a user to hit the "2" key twice for the "b" character for example. Additionally, several common subject lines and message urgency settings were provided as list selections from the GUI.

[Insert **Figure 1 approximately here. Example of the "show email" GUI display in e-PIM]**

3.2 Multimodal Synchronization

The W3C (W3C Note 8, 2003) distinguishes several types of multimodal input synchronization for input as follows.

- **sequential**: two or more input modalities are available, but only a single modality is available at any given time.
- **simultaneous**: allows input from more than one modality at the same time, but each input is acted upon separately in isolation from the others.
- **composite**: provides for the integration of input from different modes into one single request.

The e-PIM system uses simultaneous input synchronization. Each spoken or GUI submission is treated as a complete input operation by the application. A submission from the voice mode is an utterance - sometimes a request (e.g. "do I have any urgent messages from Tom") and sometimes a short reply (e.g. "no"). A submission from the GUI occurs when the user clicks on a link in the display. Input from the different modes is not currently combined into a composite request. This means for example, that the user cannot say "read me this one" and select an email on the display to have the system resolve the deictic reference.

The output synchronization is best described as form-level, which is defined by the W3C as: "all modalities are updated only at certain application defined points in the interaction" (W3C note 8, 2003). In most cases, each active modality (i.e. spoken and/or visual) is updated after each submission from the user. Delivering a synchronized result to the user proved difficult in a real deployment due to differences in latency times between the circuit-switched (voice) service and the packet-switched (data) service; for details on packet and circuit switched services see (Ruuska et. al, 2001). The system response was typically presented to the user on the voice channel slightly before it was presented visually. To reduce the impact of this problem the system always sent the visual content prior to sending the voice content. Users in the study did not observe this slight delay between the voice and the text when discussing their impressions of the system.

3.3 Device

A key requirement of the solution developed was to be completely independent of the physical client device. This was in order to eliminate the need to manage software distribution and to facilitate larger scale deployments by supporting a variety of standard devices. We piloted e-PIM on a Nokia 6650 mobile phone, which is a class A device. This device supports simultaneous circuit-switched and packet-switched connections. The device screen is full color with 128 by 160 pixels, four way scroll and with user changeable font size. The Wireless Application Protocol (WAP) browser on the device is a single threaded phone-based browser supporting Wireless Markup Language (WML) 1.3. The device supports Service Indication (SI) WAP push but not Service Loading (SL) WAP push. An SI request shows up on the device as a short text message, a URL, and a button to accept the SI. If the user accepts, the content is fetched from that URL and displayed in the WAP browser. By contrast, an SL WAP push, results in new content being displayed in the WAP browser without any user intervention. Browser support for SL was not available at the time of the pilot on the device we used. Since we did not want the interaction to require user intervention for each GUI update resulting from voice input, we used WAP push only in establishing the initial connection between the GUI browser and the application. For subsequent pushes we adopted the polling approach described in the architecture section below.

3.4 Architecture

Third generation networks support the multicall supplementary service. This capability enables concurrent connections from the mobile phone to both voice (circuit-switched) and low latency data (packet-session) networks. We leverage this capability of 3G networks to support simultaneous voice and GUI interaction. The voice network is used to establish a call between the cell phone and the speech recognition telephony server. An SMS channel in the voice network is then used to establish the initial connection between the application server and the visual browser on the phone. The data network is used to transport WML content over a WAP stack between the cell phone and the WAP gateway.

In order to support the voice interface, speech recognition and synthesis capabilities must be present either on the device or on the network. Speech embedded in the device has limited capability compared with speech resident on the network. Network-based speech, for example, supports statistical language model based recognition which can theoretically recognize any sentence constructed of words in the model's vocabulary. This type of recognition is generally used to support natural language interfaces (Jurafsky 2000). Speech embedded in the device allows recognition of only a few hundred words uttered in a structured format. Performing speech recognition on the device also places dependencies on the product development of handsets, with little guarantee of consistency of user experience across a range of handsets. As a result of these considerations, all of the speech processing for e-PIM is completed on the telephony platform server located within the network. Figure 2 shows the logical components of our solution architecture.

[Insert Figure 2 approximately here – multimodal e-PIM logical architecture]

The WAP client browser used in this e-PIM deployment was a standard phone-based browser. The initial push to the WAP browser is accomplished by sending a Service Indication message to the calling device. After that, a polling approach is used to deliver pushed information to the visual client. A URL referencing the name of the page for the next dialog turn is included in all multimodal responses prepared for the WAP browser. When the page is rendered by the WAP browser, an embedded script, within the retrieved page, issues an HTTP request to retrieve the next page when it becomes available. The push service content server, to which this request is issued, checks periodically for the requested page and returns only when it is found. In some cases, the requested document may not be found. For example, if the next user request is issued from the GUI, a normal HTTP request-response scenario ensues. With this in mind, if a new request comes in from the same application session, before an existing request is satisfied, the current request is cancelled by the push service content server. In summary, if the user makes a speech request, the resulting GUI view will be generated into a page and delivered by the content server. However if the user make a GUI request, any outstanding requests are cancelled by the content server and the next GUI page is delivered directly from the application. In either case, a new HTTP request will be issued to the content server after the system response is presented to the user.

The VoiceXML 2.0 client browser is located on a server within the network, and it is extended with two capabilities. The first is the ability to have content pushed to it asynchronously, to support multimodal interaction, and the second is support for statistical language model based recognition.

The Multicall Session Manager uses caller id and cookie management to create and maintain the association between different connections, voice and data, from the same device. This also modifies incoming requests so as to present a single client to the application manager. At runtime the content transcoder transforms the application response into a markup suitable for the client browser on which the response will be presented. We implement this by using an XSLT (eXtensible Stylesheet Language Transformation, http://www.w3.org/TR/xslt) engine and a style sheet for each supported client. Currently there are three supported clients: an enhanced VoiceXML 2.0 browser for the voice interface; a WML 1.3 browser for the mobile phone GUI implementation; and HTML 3.2 for the GUI implementation on a phone-enabled iPAQ (a Pocket PC made by Hewlett-Packard).

The application provides natural language support for the speech interface, multimodal dialog management, and the core PIM functions. In response to each request the application returns a compound document that contains content for both the voice and the visual interface.

3.5 Interaction Flows

3.5.1 Starting the application

The user starts the application by calling the phone number of the system. As e-PIM answers and commences output of the voice welcome prompt, the user is notified that he has received a Service Indication. If the user accepts the Service Indication by clicking on it (a "click" is accomplished by pressing the selection key on the phone with the item highlighted), the login page is displayed on the phone screen. At this point, the user can either say their name, or type it into the GUI browser using the telephone keypad.

3.5.2 Multimodal interaction

We did not want the voice to simply mirror the text from the visual display, and vice a versa. Since this is an NL system, the voice interface uses open ended prompts (e.g., "*what next*?") or asks for a specific piece of information related to the current task (e.g., "*what is the start time for this meeting*?"). On the other hand, the GUI presents either a list of choices or a form, with input fields, associated with the current task. So for example after the user logs in, the GUI displays the main menu and the voice prompt may say "*Where shall we start*?". When presenting the results of a user's query, the output modalities often present the same content, but sometimes only a summary is given by voice (e.g. "*you have 21 new messages*") with the details being presented visually as a list of headers on the screen.

When a user's utterance is not recognized, or when the user has been silent for too long, the voice interface provides feedback about these events (e.g., "I did not understand what you said," "I did not hear you"). In these cases, even though it is a multimodal application, we do not update the GUI and instead simply let the user try again with the original information available on the screen. In this manner, the modalities are used in a complementary fashion, rather than a redundant one. In the current implementation, the voice and text are redundant when the user either asks to have a message read to him, or selects the message body from the GUI. In either of these situations the same information is presented by both voice and text. As it turns out this redundancy was one of the aspects that users complained about in the study.

A limitation of the trial deployment, due to the fact that the version of the WAP browser was single-threaded, is that GUI input is only possible for a short window of time following a GUI refresh. When a new page is loaded, there is an application-configurable period of time during which the user can either select a link or begin text entry. A spinning globe in the upper right corner of the screen indicates that polling has begun and GUI input has been disabled. Once the globe starts to spin, users are unable to interact with the GUI until the display is refreshed or they interrupt the polling by selecting cancel. We found through experimentation prior to the start of the study that a window of 7 to 10 seconds worked best since it allows a user enough time to navigate the GUI and initiate action.

3.5.3 Ending the application

The user is responsible for managing the termination of the application. This is achieved by ending each mode independently. They can hang up the voice session and keep interacting through the GUI, or end the WAP session and proceed with a voice-only call. Only after both modalities are terminated, or a timeout period of inactivity passes, is the user session ended.

4. LABORATORY TRIAL

A user trial was conducted with a fully functioning multimodal email system in a 3G environment. The goals of the study were to:

1. Test the prototype implementation with representative users;
2. Measure the incremental value of adding natural language speech to a text-only solution;
3. Determine users' response to the multimodal system along with their willingness to purchase such a service or recommend to other users;
4. Obtain objective performance metrics for identical tasks with both a unimodal and multimodal system

In order to measure the value of the additional modality (speech), we created a baseline measurement for each participant by having him/her use a unimodal system for accessing email. The unimodal system was selected as being representative of systems that use a WAP browser for mobile mail access (see Figure 3). The same physical handset (the Nokia 6650) was used to access both systems.

[Insert Figure 3 approximately here - Representative screen for unimodal WAP browser access to email.]

4.1 Study Setup

In order to gain insight into how people used the additional modality, we had a total of 17 people use e-PIM in a lab setting. These same participants also used a fairly standard (based on the current set of available products for WAP access to email) unimodal system to perform an identical set of tasks on a similar set of messages. A within-subject design was used with each participant using both systems. The order of the mailboxes and the systems was altered to ensure that there would be no order effect, nor any effect due strictly to the messages in a particular inbox. The messages in each mailbox were balanced as to length of each message and content, such that each message in mailbox 1 had the same word count and was of approximately the same nature as the corresponding message in mailbox 2. There were 13 messages in each mailbox. Readability scores for message in mailbox 1 showed a reading ease level of 73.7, and the messages in mailbox 2 were measured at 73.2.

Participants were 17 employees (10 males and 7 females), from the three companies involved in the pilot. They volunteered for the study and had jobs that are representative of the type of user that requires mobile access to business email. The companies involved are respectively major providers of computer hardware and software, telephony services and telephone handsets. Twelve participants were in the age group 21-35, while 9 were in the age group 36-50. To avoid any potential difficulty in understanding the synthetic speech, all participants were required to be native English-speakers with no reported hearing problems. Each participant received a token gift for his/her participation.

Participants were brought in one at a time and used the same phone to access fictitious email messages. Each participant used both the unimodal system and the multimodal one with the order of the systems being counterbalanced. Participants received training on each system prior to the use of that system. Training consisted only of familiarizing them with how to log into the system and access the main menu, plus high level instructions regarding navigation on the phone and the keypad. The participants used a commercially available hands-free noise-canceling microphone with an ear bud as pictured in Figure 4. Thus, they were able to look at the screen and speak at the same time. Each task in the study could be accomplished either by GUI, by speech, or a combination of speech and GUI. When participants sent replies to email messages, a recording of the audio was included as an attachment to the email rather than sending decoded speech-to-text, due to the potential for high error rates in the decoding.

[Insert Figure 4 approximately here. Hands-free microphone with ear bud]

For each system the participant was asked to complete the following tasks.

1. Log on (account name: user one, password: 111111)
2. Find out how many messages are in your inbox.
3. Find out if you have any messages from a *named person*, if you do, read the message.
4. Send a reply to that message indicating that you are willing to cover the meeting for that person. However, it is quite possible that your calendar may not be free at that time and you should let the person know that the meeting might have to be scheduled for a different time.
5. Read the 10th message.
6. Forward that message to a *named person*, adding the following comment: “Hi David, Please see the attached message for your information.”

4.2 Measures

Performance metrics and subjective measurements of participants’ perception and attitude were measured in the study. A preliminary analysis of only the attitudinal measures was reported in Lai, 2004. Performance metrics consisted of time-on-task, number of errors and completion rates. Subjective measurements of participants’ perception and attitude were measured with questionnaires.

After use of each system, the participant completed a questionnaire consisting of attitudinal questions regarding the system, and their user experience. Participants' demographic information was collected at the end of the questionnaire. All the questions except the demographic ones were measured by asking how well certain adjectives described the system, and how the user felt while using the system with a Likert scale ("0" = "describes very poorly", "7" = "describes very well").

Composite indices were created through factor analysis to measure user experience (draining, engaged) and system perception (ease , novelty, value,). User satisfaction along with willingness to use and recommend in the future was measured on Likert scale (1= not at all, 7=very much so).

1) *Ease of Use*: consisted of "easy to use", "difficult" (reverse coded), and "straightforward"; Cronbach alpha = .823;
2) *Novelty of the system*: consisted of "outdated" (reverse coded), "cutting edge", and "innovative", Cronbach alpha = .824.
3) *Value of the system*: consisted of "useless" (reverse coded), "valuable", and "high quality", Cronbach alpha = .807.

For the user experience, an index of how draining the interaction was consisted of "exhausted", "impatient" , and "bored", (Cronbach alpha = .836). Also an engagement index was created consisting of "entertained", "involved" and "interested". (Cronbach alpha = .89)

5. FINDINGS

The results outlined below, for user perception and user performance, indicate that the multimodal interaction was significantly preferred to the unimodal access for all measurements with the exception of the ease of use index, where the difference was not significant (which was perhaps a reflection of the stability problems encountered with the 3G network where calls were dropped or would not connect). The participants were also significantly faster with the multi-modal voice system, had very high completion rates on both systems, and experienced several usability issues on each system. Table 1 presents a summary of the attitudinal findings.

5.1 User Perception

Paired sample T-tests were run to compare the means from the indices collected for each system. Participants' perception of the system's value was significantly lower for the unimodal system (M = 13.81) than the multimodal system (M = 16.93), t(15) = -2.498, p < .05. Participants also thought the multimodal system was more novel (M = 19.50) than the current mobile offering (M = 11.63), t(15) = -8.08, p < .001. Interestingly, the difference in the ease of use index was not significant, (M = 12.18 and M = 13.75) which was perhaps a reflection of the 3G network stability problems.

The user experience index confirmed the preference for the multimodal system. Participants felt significantly less drained after dealing with the multimodal system (M = 7.19) than the unimodal system (M = 11.94) t(15) = 3.288, p < .005. This was most likely due to the need to use multitap keypad input for text creation (required for email replies or forwarded comments). The multitap interface requires a user to hit the "2" key twice for the "b" character for example. However, SMS messaging is quite well established in Sydney and many, if not all, of the participants send and receive SMS messages as part of their daily business and social life. Several participants turned on the predictive T9 dictionary and achieved comfortable input rates. One user was so fluent as to use two-thumb typing on the telephone keypad. Thus clearly the entire effect of feeling more drained with the unimodal can not be attributed to the input mechanism.

Participants also had a significantly higher engagement index with the multimodal system (M = 16.75) than the unimodal system (M = 12.25) t(15) = -5.411, p < .001., and were significantly more likely use and recommend the multimodal system in the future (M = 6.31) than the unimodal system (M = 3.94) t(15) = 5.69, p < .001.

5.2 User Performance

The total task time was significantly faster (37%) for the multimodal system (M=9.49) t(15) = 2.56, p < .05 than for the unimodal system (M=12.98 minutes). If we remove the three participants whose times were increased due to getting stuck in a known e-PIM navigation problem, the difference is more pronounced (50% faster) and still significant (M=13.44 for unimodal and M= 8.96 for e-PIM) t(12) = 2.89, p < .05. If we further remove a participant whose quips and wise cracks (e.g. in response to a yes or no question the participant replied "of course I don't want to delete it you stupid machine") greatly confused the speech recognition engine, the difference in the mean times is greater still (63% faster) and continues to be significant (M=14.05 for unimodal and M= 8.61 for e-PIM) t(11) = 4.12, p < .005. Figure 5 charts the means for these times (in minutes).

Table 1. Summary of significant attitudinal findings

Index	**uni**	**multi**	**Significance**
Drained (max=21)	11.94	7.19	p < .005
Engaged (max=21)	12.25	16.75	p < .001
Value (max=21)	13.81	16.93	p < .05
Novel (max=21)	11.63	19.50	p < .001
Ease of Use (max=21)	12.18	13.75	p < .05
Likely to use (max=7)	3.94	6.31	p < .001
Likely to recommend (max=7)	4.0	6.38	p < .001
Satisfied with amount of time required (max=7)	2.56	4.75	p < .005

[insert Figure 5 approximately here. Means for total times in minutes]

5.3 Interaction Metrics

There were 29 complete recognition failures out of a total of 293 speech turns; in addition, there were 32 GUI turns. Of the recognition failures, 26 were recovered by the user, while on three occasions the user decided to hang-up and terminate the interaction (excluded are partial recognition failures, which may be successfully modeled by the NLU). The following diagram, Figure 6 , illustrates the distribution of these recognition failures in the context of the proportion of speech and GUI turns per user.

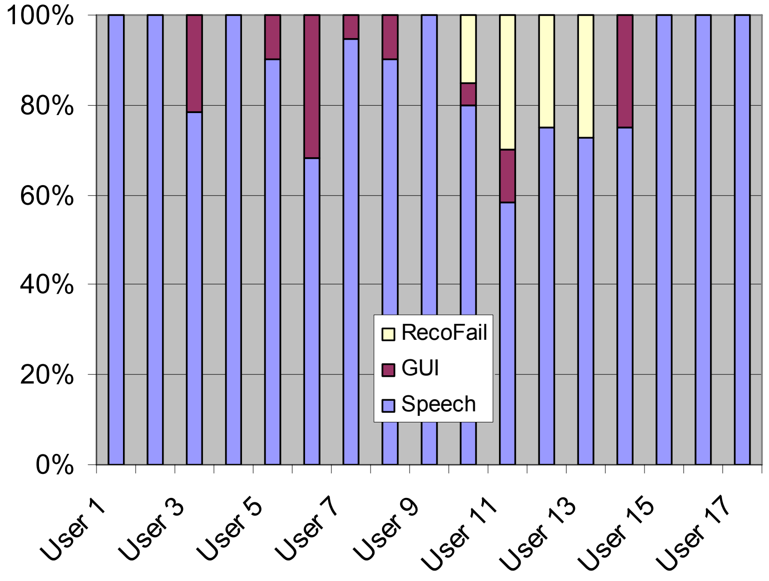


Figure 6. Proportion of GUI, Speech turns and Recognition failures.

Time per task varied considerably amongst the trial users, ranging from five seconds up to several minutes. There was a high completion rate observed, with 86 out of 102 (17 users by 6) tasks completed successfully. The 16 tasks that were incomplete included those tasks which were only partially completed. The following diagram, Figure 7, depicts the average time spent per task by each user for each of the trial tasks.

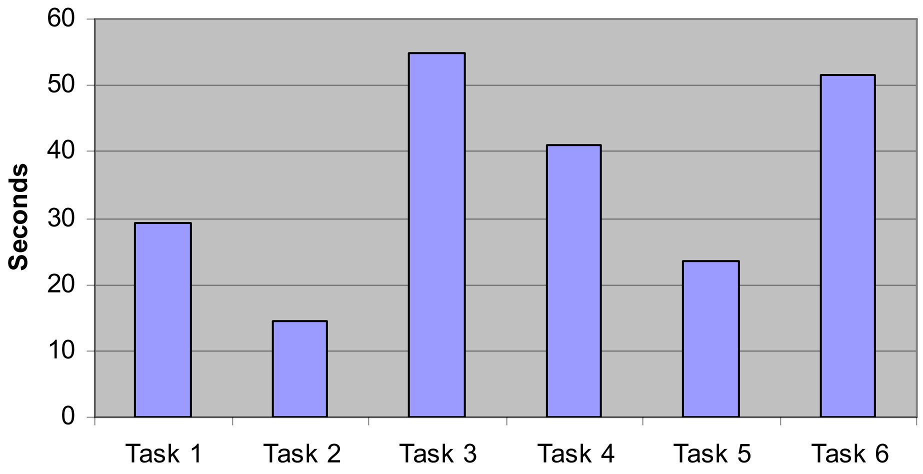


Figure 7. Average duration of time per task

6. ANALYSIS OF DESIGN

6.1 Modality

In addition to examining the incremental value of adding speech to a graphical interface in a mobile setting, the user study contributed to increasing the understanding of modality usage (i.e. which modality do users prefer given the task and the circumstances). Our hypothesis was that there would be a distribution of usage between speech and GUI for navigation depending on the preferences of the participant and his/her rate of success. However, this was not the case. Speech was used the majority of the time for both input and navigation even though most of the participants had little to no prior experience with speech systems and did have prior experience with text-only access to applications on their phones. Speech was the dominant modality used in the multimodal system (both for navigation and input), accounting for 90% of all turns. The following diagram (Figure 8) illustrates for each participant whether speech or data (via the GUI) was used in a given turn; a more detailed analysis of this result is outlined in (Pavlovski et. al, 2004).

[Insert Figure 8 approximately here - Speech versus GUI usage in a given turn.

Part of the explanation for the dominance of the use of the speech modality may have been the novelty of the interface for these users (note the high ratings for Novelty in Table 1) however it is generally not the case that new technologies are rapidly adopted by all new users. Another explanation is that WAP on a cell phone is truly an impoverished GUI. Given the choice between using a reasonable speech interface or inputting text with a telephone keypad and a tiny screen, it is not surprising that most users selected speech for input. However we also saw participants using speech for navigation even under circumstances where they were not being well understood (e.g. high levels of background noise). Due the presence of the noise-canceling microphone, most users had fairly high recognition rates (recognition rates appeared to be above 90% for most users/tasks). The one exception to the high accuracy was when users were instructed to listen to a specific message in task 5 ("Read the 10th message"). The numeric ("10th") used in the utterance often caused recognition errors. A final explanation for the predominant use of speech in the multimodal system is speed of response time. One participant commented "*It's just easier. It's quicker. And the response time is so quick, it speaks back so quickly. I think if you had to wait a long time for the response that you would probably find it quicker with the keyboard*." It is not clear that this modality dominance would hold up over time. We would expect that as users had time to familiarize themselves with both interfaces they would move fluidly between modalities depending on the circumstances of use.

A further hypothesis of the study was that participants would fall back rather quickly to GUI usage when speech recognition failed. This hypothesis was based on prior research (Oviatt, 2003). Thus we were surprised to find that in most cases users persevered with speech, trying not only to alternate phrasings, but returning to phrasings that had previously failed. A representative comment was "*It's almost like a learning curve thing… I guess that is why I went back to trying different variants. If you can get that (speech) working once as a user, you've got that problem solved forever. It's like an investment in that interface rather than falling back to the screen.*"

A further consideration to the above results is that since the sending of emails is conducted by recording a message first, it is not straightforward to compare user behaviour directly with the unimodal system. Although, in the context of multimodal usage the user does have the option of completing the send message task with either a speech or GUI command, after initially recording their message.

6.2 Fusion of Inbound Requests

While WAP may be an impoverished UI, it is representative of mobile email applications available today, and the value of adding speech has not been clearly shown to date. Our prior experience with speech-only applications in a mobile environment showed that providing speech-only email access is problematic for many users given the need to navigate through long lists of new messages, complicated messages bodies (e.g. embedded tables in the message) and attachments. In the multimodal version of e-PIM tested during this laboratory trial, the output to screen for visual scanning adds richness to the interface that can not be discounted when evaluating the benefit of the multimodal system. While it is perhaps not surprising that users preferred the availability of both speech and GUI to GUI-only, it was surprising that users chose speech over GUI selection even in noisy environments and even when speech was not working well for them.

One could debate whether e-PIM is technically a true multimodal system or not because there was no "fusion" of data coming from different input streams and thus no ability to use mutual disambiguation. Mutual disambiguation refers to the ability to increase recognition accuracy by integrating input in such a way that each mode provides context for interpreting the other during integration (Oviatt, 1999). However, while the system does not currently support the deictic reference "play this message" (voice plus selection), modalities can be mixed within a given task (e.g. sending an email message) as in the example below.

User: *send email (incomplete utterance),*

System: *who is this message going to?*

User: *Jennifer Smith*

System: *Sorry?*

User: *User keys in name*.

We believe the more interesting discussion relates to the need to define guidelines for designing multimodal systems in such a way that the modalities are used in a complementary fashion and the strengths of each modality can be capitalized upon by the user depending on preferences and context of use. This discussion must cover the input modalities, (which is already being investigated in the areas of speech and pen gesture as well as speech and lip reading), but also ways of achieving effective cross modal integration for output.

9. DESIGN IMPLICATIONS

A usability issue frequently mentioned by participants using e-PIM was not knowing what can be said to the system. When users are presented with a task (e.g. browsing email messages) they want to take an action but are unsure of the words to speak to accomplish this. The major benefit to using a natural language understanding system is that there are no "correct" or "incorrect" phrasings, as there are in grammar-based speech systems. However, there are "supported" and "unsupported" functions and users were occasionally unsure whether they had inadvertently wandered into an unsupported function. Thus when a user says "*do I have any messages about the special seminar*" and the system replies: "*I'm sorry I don't understand*" it is unclear to the user whether this is because a key word search is not supported, or that particular utterance was not understood. This problem is common to all speech systems and there are guidelines for ameliorating the problem (Lai 2002, Cohen et. al, 2004). Given additional development time, the spoken interface could have been modified to support these guidelines.

Design implication: Even though the speech interface is supported by the visual output to the screen, the issue of invisibility (Lai 2002) remains. Thus the speech side of the multimodal interface must follow conversational design guidelines to help the user know what to say.

More interestingly, a usability problem that surfaced which is specific to multimodal systems is a feeling of "overload" when both modalities are presenting at the same time. Also note that while feeling "drained" (see Table 1) was significantly higher with the WAP text-only system, it was not extremely low either for the multimodal system. As mentioned earlier, when the system presents an email message the text of the message is displayed on the screen, and the text is "spoken" by the system using text-to-speech. We presented the information in both modalities for the body of the message because we don't know if the user is in an eyes-free situation such as driving, or is sitting someplace where he can easily read the text. Given the divided-attention state of the user in a mobile setting, we used complementary output modes for navigation feedback, but used redundant modes when presenting important text such as the body of the email. One participant commented: "*When I pressed the select key, it started reading it (the message) out to me again and I started to feel overwhelmed. I found it very confusing swapping between the modes. One of the things that was adding to the confusion is that there was too much information at once."*

Design implication: Just as the input modes to a multimodal system should ideally be complementary, the output modes should reinforce each other without being redundant. Participant feedback indicates that audio combined with redundant text is not effective in this context for increasing user comprehension.

The voice system was enabled with barge-in, which allows a user to interrupt the system while it is speaking. Many users intuitively tried this function - sometimes by jokingly telling the system to "shut up", but more often with the term "stop", which worked well. Barge-in however only causes the system to stop speaking its current utterance, and to listen for the next command, usually with an open-end prompt such as "How can I help?" If no reply is received after a while, the system will re-prompt with something along the lines of "I'm sorry, I didn't hear anything." Thus if the user is finding the spoken speech to be distracting, barge-in is not a sufficient remedy. The inclusion of a mute function to suppress spoken output was mentioned by several users as a helpful function to include. A representative comment was: *"Can there please be a command for shut up. Because I was starting to get really frustrated when it was speaking and I was trying to navigate through it."*

Design implication: The system should support a spoken command for "be quiet" and not resume speaking until it is asked to do so. It should also have a way to mute the speech using the GUI.

Users commented that when interacting with the system if another person asked a question (i.e. interrupted them while on the phone), to which they responded, the system is not able to discern commands to interpret versus conversations with other people. Furthermore, the system may sometimes respond to the other person's questions if they were within close proximity. (This appeared to have occurred on rare occasions and was observed infrequently.)

Design implication: There may be a need to support a name or identifier for the system, so that the system is able to detect specific commands for it to act upon.

When the GUI was polling the server it displayed a globe spinning in the upper right-hand corner of the screen. This was our feedback of "GUI busy" to the user, and it indicated that the GUI was not available for interaction. This concept should not have been unfamiliar to users who have used the internet since it is similar to the busy indicator available in standard web browsers (e.g. flag waving in the upper-right hand corner of Internet Explorer). As mentioned earlier, if the user wanted to input text or make a GUI selection while the globe was spinning, he would have to first press Cancel, to interrupt the polling operation. This was explained to users during the training period prior to the start of the study. However we found that during the study, many users missed this indication and tried to use the GUI even if it was busy. One participant commented: *"When it wasn't getting me, I tried to move to the graphical one and some of that was not working too well, when I pressed the select key."* When users fell-back to GUI usage, it was usually because they had encountered multiple problems with speech on a given task, and thus they were probably somewhat flustered at that point.

Design implication: Given that we know multimodal users in a mobile setting are most likely in a divided-attention state, and the screen on a mobile phone is small and subject to sun glare, visual feedback of system state needs to be highly noticeable.

Additionally, there were many usability issues that participants encountered with the unimodal system. Since the focus of the paper is on the multimodal system, we won't devote much space to these issues but simply mention that text input was a major problem for all participants, even those that were clearly adept at text messaging and using T9 predictive input. Participants also had problems entering the text in the wrong area (e.g., entering the message reply in the area that was dedicated to entering the recipient's address), finding the necessary function (a lot of time was spent looking in vain for Reply in the Options menu) and getting a feeling for the size and scope of messages in their inbox (for example, they could not tell without scrolling through 4 or 5 screens how many new messages had been received).

10. CONCLUSIONS

In this paper we present a description of the implementation of e-PIM, a conversational multimodal application for mobile email retrieval with multimedia output (synthetic speech and text). E-PIM accepts free-form natural speech, GUI selection and keyed text for input. The system was successfully used in a laboratory trial on an unmodified cell phone in a 3G environment. Modifying the phone would have made the development much easier, but would have limited the distribution and availability of the solution. We describe the architecture used for the implementation of e-PIM and highlight issues encountered with the real-world implementation of the solution (e.g. data channel is slower than voice channel, lack of browser support for service loading) along with proposed solutions.

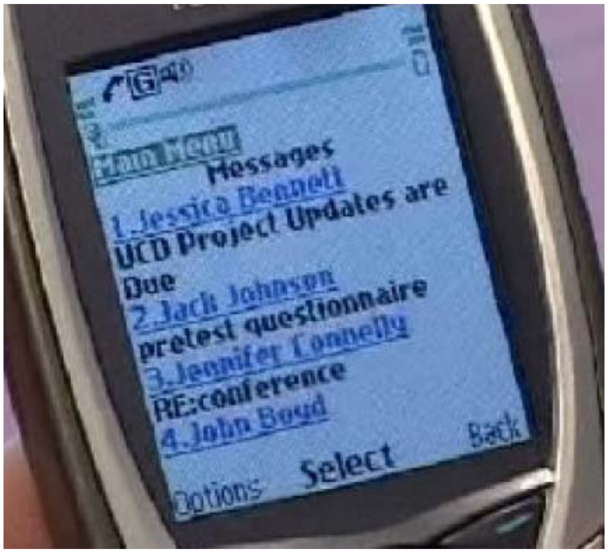
The paper also presents results from a laboratory trial of the e-PIM multimodal application. The user trial consisted of 17 users. Our findings indicate that participants were significantly faster with voice and text than with text alone, and that the multimodal system was significantly preferred by users. We also discuss usability problems and modality issues observed during the trial. The associated design implications from these issues are highlighted to help inform future designers of multimodal systems for mobile applications.

More work needs to be done to create a similar system that can accept blended input, which would both and could possibly create higher recognition rates through use of mutual disambiguation.

11. ACKNOWLEDGEMENTS

We would like to thank our technical colleagues on the e-PIM team in New York as well as in Asia Pacific for making the solution and the user trial move from the domain of sketches on paper to reality.

12. FIGURES and TABLES



Messages

1.Jessica Bennett
UCD Project Updates are Due

2.Jack Johnson
pretest questionnaire

3.Jennifer Connelly
RE:conference

4.John Boyd

Figure 1 - Example of the “show email” GUI display in e-PIM

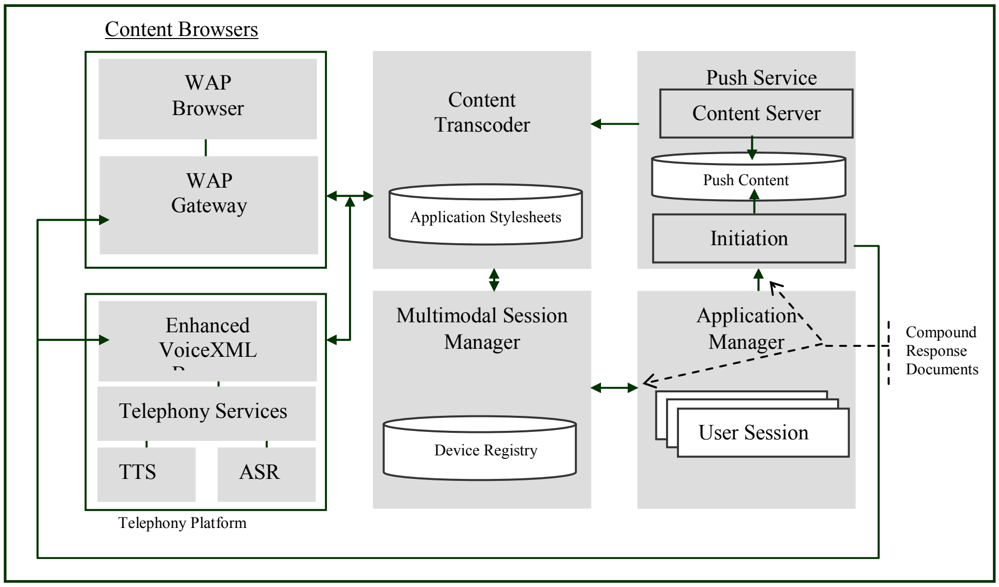
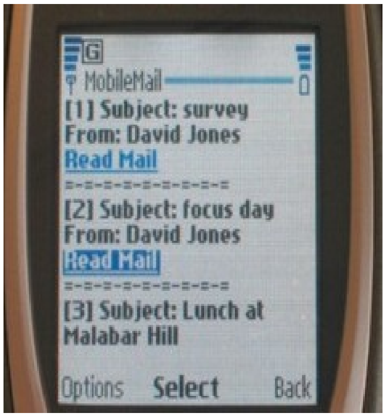


Figure 2 – Multimodal e-PIM logical architecture



MobileMail

[1] Subject: survey
From: David Jones
Read Mail

=-=-=-=-=-=-=-=-=

[2] Subject: focus day
From: David Jones
Read Mail

=-=-=-=-=-=-=-=-=

[3] Subject: Lunch at
Malabar Hill

Figure 3 - Representative screen for unimodal WAP browser access to email.

Figure 4 - Hands-free microphone with ear bud

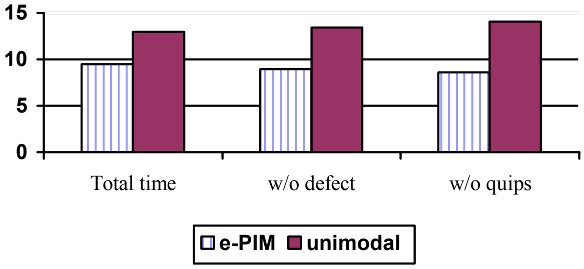
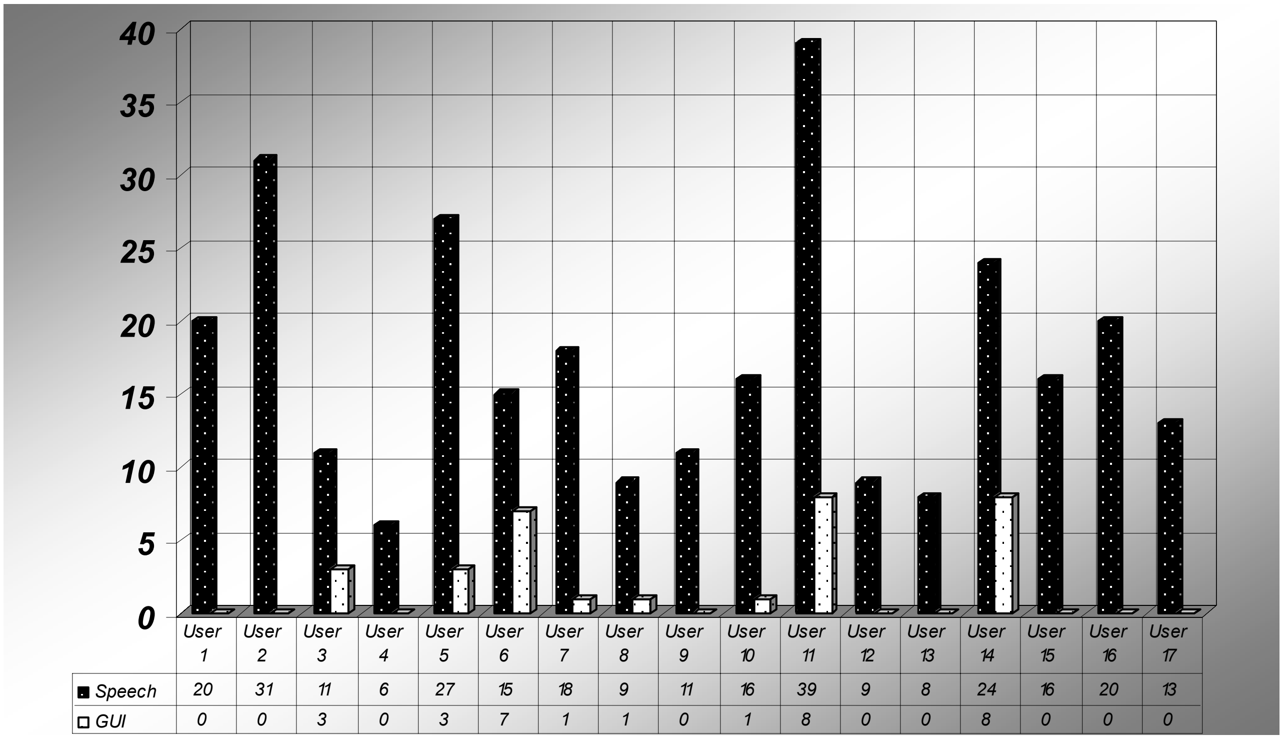


Figure 5 - Means for total times in minutes



	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9	User 10	User 11	User 12	User 13	User 14	User 15	User 16	User 17
■ Speech	20	31	11	6	27	15	18	9	11	16	39	9	8	24	16	20	13
□ GUI	0	0	3	0	3	7	1	1	0	1	8	0	0	8	0	0	0

Figure 8. Speech versus GUI usage in a given turn

13. REFERENCES

1. Baddeley, A. (1992). Working Memory. *Science* vol. 255, pp. 556-559.

2. Cohen, M., Giangola, J., Balogh, J. (February 2004). *Voice User Interface Design.* Addison-Wesley Professional.

3 Cohen, P., McGee, D., Clow, J. (April 2000). The Efficiency of Multimodal Interactions for a Map-based Task. In *Proceedings of the sixth conference on Applied natural language processing, pp. 331-338.*

4. Gong, L. (2003). Multimodal interactions on mobile devices and users' behavioral and attitudinal preferences. In C. Stephanidis (Ed.), Universal access in HCI: Inclusive design in the information society (pp. 1402-1406). Mahwah, NJ: Lawrence Erlbaum Associates

5. Grover, D.L., King, M.T., and Kushler, C. A. (1998). Patent No. US5818437. Reduced keyboard disambiguating computer. Tegic Communications, Inc., Seattle.

6. James, C., Reischel, K. (2001). Text input for mobile devices: comparing model prediction to actual performance. In *Proceedings of the Conference on Human Factors in Computing Systems, pp. 365-372.*

7. Jurafsky, D., Martin, J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall.

8. Lai, J., Mitchell, S., Viveros, M., Wood, D., Lee, K.M. (2002). Ubiquitous Access to Unified Messaging: A study of Usability, and the Limits of Pervasive Computing. *International Journal of Human Computer Interaction,* Volume 14 (3-4).

9. Lai, J., Yankelovich, N. (2002). Conversational Speech Interfaces. In *The Handbook of Human Computer Interaction*. Sears, A., Jacko, J. (Editors) LEA, New Jersey.

10. Lai, J. (2004), Facilitating Mobile Communication with Multimodal Access to Email Messages on a Cell Phone, In *Proceedings of the Conference on Human Factors in Computing Systems – Late Breaking , CHI 2004*

11. Longueuil, D. (2002). Wireless Messaging Demystified: SMS, EMS, MMS, IM, and other. McGraw-Hill Professional.

12. Mayer, R. E. & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual information processing systems In working memory. *Journal of Educational Psychology, 90,* 312-320.

13. Oviatt, S. (1996). Multimodal interfaces for Dynamic Interactive Maps. In *Proceedings of the Conference on Human Factors in Computing Systems, pp. 95-103. CHI '96.*

14. Oviatt, S. (1999). Mutual Disambiguation of Recognition Errors in a Multimodal Architecgture. In Proceedings of the Conference on Human Factors in Computing Systems, CHI '99, NY, pp. 576-583.

15. Oviatt, S. L. (2000). Multimodal Signal Processing in Naturalistic Noisy Environments. In Proceedings of the International Conference of Spoken Language Processing (ICSLP '2000), ed. By B. Yuan, T. Huang & X. Tang. Beijing: Chinese Friendship Publishers. Vol 2, pp. 696-699.

16. Oviatt, S.L. (2002). Breaking the Robustness Barrier: Recent Progress on the Design of Robust Multimodal Systems. In *Advances in Computers* (ed. by M. Zelkowitz). Academic Press, vol. 56, 305-341.

17. Oviatt,S. (2003). Multimodal interfaces. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, (ed. by J. Jacko and A. Sears). Lawrence Erlbaum Assoc., Mahwah, NJ, chap.14, 286-304.

18. Pavlovski, C., Lai, J., Mitchell, S. (2004). Etiology of User Experience with Natural Language Speech, ICSLP 2004.

19. Ruuska, P., Frantti, T., (September 2001). The Multicall Service to Support Multimedia Services in the UMTS Networks. In Proceedings of the 27th Euromicro Conference, Warsaw, Poland.

20. Salonisdis, T., Digalakis, V. (1998). Robust Speech Recognition for Multiple Topological Scenarios of the GSM Mobile Phone System. International Conference in Acoustics and Signal Processing (ICASSP), May 12-15 1998, Seattle, USA.

21. Sawhney, N., Schmandt, C. (2000) Nomadic Radio, ACM Transactions on Computer-Human Interaction. Volume 7, No. 3.

22. Wickens, C., Sandry, D.,Vidulich, M. (1983). Compatibility and resource competition between modalities of input, central processing, and output. In Human Factors, vol. 25, pp. 227-248.

23. W3C Note 8. (January 2003). Multimodal Interaction Requirements http://www.w3.org/TR/mmireqs

24. Yankelovich, N., Levow, G., Marx, M. (1995). Designing SpeechActs: issues in speech user interfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems*, p.369-376, May 07-11, Denver, Colorado, United States.

.