Facilitating Mobile Communication with Multimodal Access to Email Messages on a Cell Phone

Jennifer Lai
IBM T.J. Watson Research Center
Hawthorne, NY 10532
jlai@us.ibm.com

ABSTRACT

This paper reports on a user trial (N=17) that compares the use of two systems for accessing email messages on a telephone handset. The first system uses graphic output and telephone keypad input, while the second system has both graphic and speech output, with keypad and speech as input. To our knowledge, this trial represents the first evaluation of a fully functioning multimodal system that uses natural language understanding on a phone, and was dependent on the 3G network currently available in Australia. Participants saw significantly greater value in the multimodal interaction, and rated their experience with the multimodal system significantly more positively than the unimodal system. They were also significantly more inclined to use and recommend the multimodal system over the current unimodal product offering. While we expected to see some mixed usage of modalities in the multimodal system, participants used speech predominantly, falling back to GUI selection only after encountering multiple speech recognition failures in a row.

Author Keywords

Multimodal interaction, mobile interface design, pervasive computing, mobile communication, spoken interfaces

ACM Classification Keywords

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

INTRODUCTION AND MOTIVATION

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 15 to 20 years [3]. They are often viewed as the solution to increasing the robustness and accuracy of speech-only systems, and are particularly well suited for use in a mobile computing environment given the varying constraints placed on both the user and the recognition technology [2]. The additional mode can be used either in a complementary way (to supplement the recognition technology), in a redundant manner, or as an alternate to the recognition technology in case of high error rates. Multimodal systems often combine more than one form of output as well, creating multimedia effects by using visual and auditory output. While most systems today are only capable of processing two modes of input, there is research in the biometrics field that is showing good success combining three or more inputs including voice, handwriting, fingerprints or retinal scans for identity verification in challenging conditions [3].

Creating a highly usable interface for browsing and accessing information over a 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 mobile email messaging systems today, does not work well when applied to large amounts of textual information. When viewing the daily onslaught of messages in a GUI setting we routinely do a visual triage, scanning for messages from people who are important to us, or for message topics that pique our interest. This triage is difficult to do when relying on auditory input, which is slower than the visual channel. A multimodal interaction model appears to be well 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. While intuitively we felt that multimodal access would be preferred by users to unimodal access, we wanted several concrete measurements of the differences. We were also interested in the combination of modalities and understanding the *what*, *how* and *when* of users' selection of modality: what modality was preferred for which task, how did users respond to the interplay of modalities, and when was one modality preferred over another. In order to measure the differentiation of the additional modality/channel (speech), we chose to baseline the study by having all participants use both a unimodal and multimodal system for accessing email over a telephone with a WAP browser.

This paper presents the findings from a user study that compares the current product offering for mobile email access from a major telephony service provider in Australia, to a multimodal prototype developed by a

CHI 2004, April 24–29, 2004, Vienna, Austria.
ACM 1-58113-703-6/04/0004.

research lab in the United States. The trial was conducted in Sydney since it required a third generation (3G) network in order to transmit voice and data at the same time. To our knowledge this represents the first study of an implemented multimodal system (speech and text) on a telephone handset. Existing products for mobile email access provide unimodal access, either voice only, or GUI only.

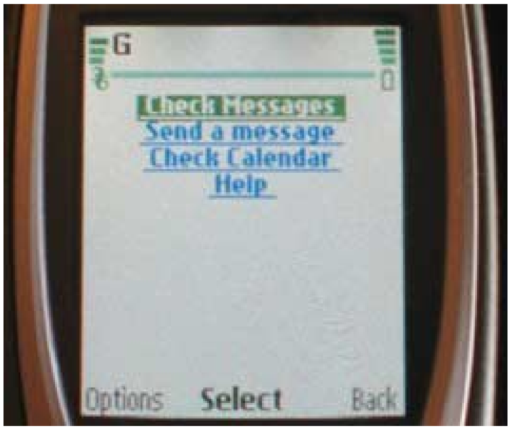


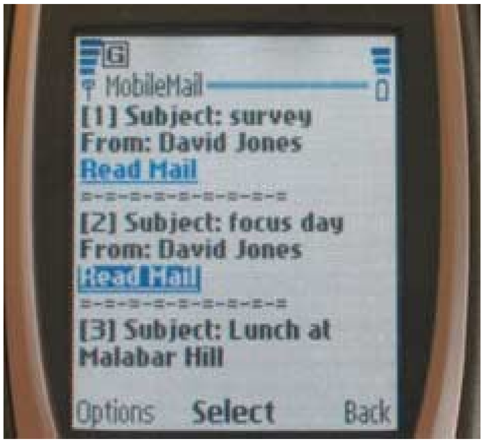
Figure 1. A screen shot from Mobile Assistant

THE SYSTEMS

The Mobile Assistant project looks at the challenge of telephone access to textual information in the context of a mobile message retrieval application. The *mobile assistant* [1] can read email messages, book appointments, take messages, and provide access to address book information. Key components are a conversational interface using speech recognition, synthetic speech, a contextual dialog engine, and notifications tailored to user preferences. The focus of the research has been on supporting the pressing communication needs of mobile workers and overcoming technological hurdles such as high accuracy speech recognition in noisy environments, natural language understanding and optimal message presentation on a variety of devices and modalities.

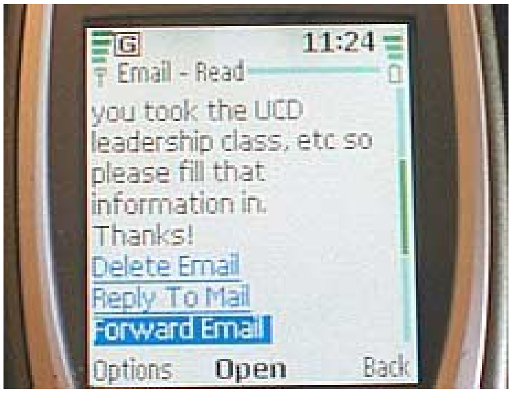
The most recent implementation of the mobile assistant moves the system from a speech-only interface to a multimodal interface by incorporating the additional modalities of graphic input/output through a WAP browser and telephone keypad, to the existing modalities of speech recognition and speech synthesis. Adjustments were made to the acoustic and language models to conform to the Australian accent and phraseology. Given that a synthetic speech engine with an Australian accent was not available, an English (UK) accent was used. The GUI interface (see Figure 1) was not unlike the graphical interface available for the unimodal system since is dependent on what the WAP browser will support.

The Mobile Assistant was compared to a unimodal system that is representative of current WAP browser clients for delivering mobile access to email messages. Message headers are listed in sets of four messages (see Figure 2), and once a message is read, the reply, forward and delete functions are available (see Figure 3). The unimodal system is a current product offering by a major telephony service provider in Australia. The actual name of the product can not be disclosed at this time due to confidentiality issues associated with the pilot.



[1] Subject: survey
From: David Jones
Read Mail
=-=-=-=-=-=-=-=-=
[2] Subject: focus day
From: David Jones
Read Mail
=-=-=-=-=-=-=-=-=
[3] Subject: Lunch at
Malabar Hill

Figure 2. The unimodal email client listing message headers



you took the UCD leadership class, etc so please fill that information in. Thanks!

Figure 3. The unimodal system with the forward action selected at the bottom of an email message.

PILOT STUDY

In order to gain insight into how people used the additional modality, we had a total of 17 people use the Mobile Assistant 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.

Experimental design

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 (see Table 1). The messages in each mailbox were balanced as to length of each message and content, such that the 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 (see Table 2 for readability scores of each mailbox)

User 1	System 1 / mailbox 1	System 2 / mailbox 2
User 2	System 2 / mailbox 1	System 1 / mailbox 2
User 3	System 1 / mailbox 2	System 2 / mailbox 1
User 4	System 2 / mailbox 2	System 1 / mailbox 1
User 5	Pattern repeats	Pattern repeats

Table 1. Order of systems and mailboxes tested

	Flesch Kincaid	Reading Ease	Reading Level
Mailbox 1	6.2	73.7	7.6
Mailbox 2	6.1	73.2	7.7

Table 2. Flesch readability and ease scores for both mailboxes

Participants

Participants were 17 employees (10 males and 7 females), from the three companies involved in the pilot. The companies are each respectively major providers of computer hardware and software, telephony services and telephone handsets. Initially 20 participants were scheduled, (balanced for gender), however due to instability with the 3G network, in the end only 17 subjects were run. 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 native English-speakers with no reported hearing problems. Participants received a token gift for their participation.

Tasks

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

1. Log on (test 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 Denise Richards, 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 9th message.
6. Forward that message to David Jones, adding the following comment: “Hi David, Please see the attached message for your information.”

Measurements

Two types of measurements were used in the study. First, a behavioral metric of participants’ task performance as measured by time-on-task, number of errors and completion rates. Secondly, subjective measurements of participants’ perception and attitude were measured with questionnaires. This paper will report only on the latter, since the time-on-task for each task is not yet available for all tasks/all participants since several timings still need to be checked by reviewing the videotapes.

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 on a Likert scale (“0” = “describes very poorly”, “7” = “describes very well”).

Three system indices were constructed through factor analysis:

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

RESULTS

For the results presented below, only the data from 16 participants was used since one of the study participants was not able to complete the unimodal session at the lab due to a server outage. Thus, while we allowed him to use the multimodal system since his curiosity was piqued; we were not able to include his data.

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, $\underline{p}$ < .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, $\underline{p}$ < .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 stability problems that we encountered with the 3G network where calls would be dropped or not connect. Figure 4 presents the mean differences for the three indices.

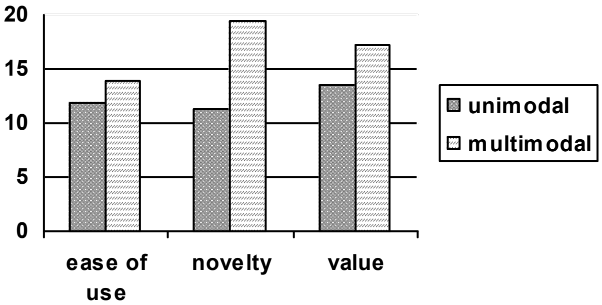


Figure 4. Comparing means for the ease, novelty and value indices for Mobile Assistant and the unimodal system

The user experience indices confirmed the preference for the multimodal system. Participants felt significantly less drained after dealing with Mobile Assistant (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 (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.

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., 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.

Speech recognition systems are not as widely available in Australia as they are in the United States, where it now seems that many of the customer care centers are running speech systems. This difference may have contributed to the high novelty ratings and high engagement factor received by the multimodal system.

Modality Usage

Given that every task in the multimodal system could be completed either with speech or with GUI usage, we expected to see some distribution between these two modalities. We were curious if a dominant modality would emerge for any given task, and what circumstances would cause users to switch modalities.

Speech input was the dominant modality used for all of the tasks presented. Perhaps in part due to the "speech technology naiveté" of the participants, users started with speech and continued throughout the session. Even one user, who had told the experimenter before the start of the test that he would use the graphical interface exclusively, (because that is the modality that he was "most familiar and comfortable with"), ended up using only speech in his session. When queried about this modality choice after the session, he replied: "I wanted to try using voice because that is what I think most people will use. And once I started using it, I just kept going with it."

When users did switch to GUI selection for navigation (none of the participants used the GUI for text input) it was always due to the presence of repeated speech recognition errors. Interestingly, the users' first several choices when encountering a speech recognition failure was to rephrase the request, or to repeat it, rather than falling back on GUI selection immediately. This finding was counter to our expectation that some users would use GUI navigation simply due to personal preferences, and that all users would quickly fall back to GUI selection when encountering speech recognition problems.

CONCLUSION

We have presented a study of a multimodal system that combines speech technologies with a standard graphic interface for facilitating access to email in a mobile communications setting. This system was compared to a current product offering for mobile access to email messages. The multimodal system was found to provide significantly more value to users when compared to the text-only system.

While these findings are perhaps not surprising, they are a confirmation of the value of pursuing multimodal designs for mobile applications. Many of the findings to-date for multimodal speech and text systems have been conducted with simulation systems, where the speech understanding portion of the system is simulated using a wizard of oz setup.

ACKNOWLEDGMENTS

This paper would not have been possible without the participants in our study, our colleagues at IBM Australia and all of the members of the Mobile Assistant group.

REFERENCES

1. Lai, J., Mitchell, S., Viveros, M., Wood, D., Lee, K.M., 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) 2002
2. Oviatt, S., Ten myths of multimodal interaction, *Communications of the ACM,* Vol. 42, No. 11, November 1999, pp. 74-81
3. Oviatt, S.L. 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, 2003, chap.14, 286-304.