Choosing Speech or Touchtone Modality for Navigation within a Telephony Natural Language System

Jennifer C. Lai

IBM T.J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532
lai@watson.ibm.com

Kwan Min Lee

Annenberg School for Communication
University of Southern California
Los Angeles, California 90089
kmlee@stanford.edu

ABSTRACT

In this paper, we describe the empirical findings from a user study (N=16) that compares the use of touchtone and speech modalities to navigate within a telephone-based message retrieval system. Unlike previous studies comparing these two modalities, the speech system used was a working natural language system. Results indicate that a majority of users preferred interacting by speech even though the touchtone interface was more efficient as measured by the success rate and speed for each task. The interaction with the speech modality was rated as being more satisfying, more entertaining, and more natural than the touchtone modality. This is an interesting demonstration of how user satisfaction may not always correspond completely with efficiency measures in task-based scenarios. The findings underline the importance of examining speech interfaces from other perspectives in addition to efficiency maximization. The paper also describes a speech-based telephony application, which is used for access to unified messaging (email, voicemail, and faxes) and calendar information. The application is currently deployed at IBM Research and in use by more than 200 users.

1. INTRODUCTION

As Voice User Interfaces become more common, users occasionally experience the discomfort of speaking to a computer by phone, in a public setting. Repetition of the same request in response to recognition errors and the use of simplistic commands can cause nearby observers to stare, or even eavesdrop. If these systems are to be used in public, as the pervasive computing paradigm dictates they should, application designers will want to accommodate users’ need for privacy and desire to be considerate to other people. Additionally, users who use a cell phone to call a speech-based system are subject to the vagaries of varying levels of cell coverage. When cell signal is low, there is a large and negative impact on the accuracy of the speech recognition, making the system unusable with speech.

Prior to the more generalized availability of speech recognition software and development toolkits, most telephone-based interfaces relied on touchtone input, often referred to as Dual Tone Multiple Frequency (DTMF) input. Voice processing systems, which first started to appear in the 1980s, used DTMF for input and recorded human speech as output [1]. The fixed set of twelve keys (ten digits as well as the # and * keys) on the keypad lent itself to the construction of applications that present the caller with lists of options (e.g., “press one for Sales”), commonly referred to as menus. Since that time, menu-driven Interactive Voice Response applications have become a pervasive phenomenon in the United States [6]. In many ways, the adoption of speech-based telephony interfaces has occurred as a move away from the restrictions of DTMF input. Speech is a powerful input mechanism, which is far superior to DTMF for entering text, and gives users direct access to specific functions without going through multiple layers of menus.

DTMF input has the significant advantage of being silent, instantaneous, and 100% accurate. This is in contrast to speech recognition, which is none of these. Processing of the speech can cause delays for the caller, referred to as latency, and variability of the spoken input prevents speech recognition from being 100% accurate. However, given the limitation of using only twelve keys for input, many users of Interactive Voice Response systems report feeling handcuffed and constrained, often causing them to hang up in frustration.

Given the tradeoff between using natural input that can be erroneously interpreted (speech) or highly constrained but consistently accurate input (DTMF), we wondered which modality users of a telephone-based message retrieval system would prefer for navigation.

This question has been pondered before. Earlier studies [2,5,8] have compared speech input with DTMF. While these studies showed that touchtone input was preferred to single-word command & control systems, the results were inconsistent and have to be viewed in light of the speech systems that were used at the time. Delogu et al [2] did not use a real speech system, but tested instead with a wizard-of-oz implementation, only simulating real speech recognition. They found no difference in terms of task completion time, and number of turns per task when comparing DTMF to speech input for an Interactive Voice Response system.

In a study reported by Karis, [8] thirty-two subjects performed call management tasks twice, once using speech and once with touchtone. The tasks involved interacting with a "single telephone number" service that included a variety of call management features. The majority of subjects (58.1%) preferred using DTMF rather than speech, and tasks were completed faster using touchtone.
We opted to repose the DTMF vs. voice question using a Natural Language system. (For a description of the NL implementation used in the system see [3]). This system uses statistical based language model recognition for the speech-to-text decoding, and transforms the utterances into a formal language representation. The formal language is passed to the dialog system, which operates on the requests and generates the system output. The system output consists of text-to-speech (TTS).

2. THE MOBILE ASSISTANT

This experiment was conducted within the framework of a working application called the Mobile Assistant (MA). This system gives users ubiquitous access to unified messages (email, voice mail and faxes) and calendar information from any telephone. Calendar and messages can be accessed in one of three modalities:

1. From a desktop computer using a combination of audio and visual modalities. Voicemail is received as an audio attachment and can be listened to within the email application inbox. For internal calls, the identity of the caller is listed in the header information of the message. Email messages can be sent using the MA system and are also received as audio attachments.

2. From a SmartPhone (a cell phone with a multi-line display and a web browser) in a silent visual mode. In this case, users connect over the network and can read their email and calendar entries on the phone’s display. Notifications of the arrival of urgent email messages and voicemail are usually sent to this phone, however the user can tailor what messages they get notified of and which email-addressable device they want the notifications sent to. Voicemail messages cannot be read on the screen since we do not transcribe them and thus they must be accessed by calling in to the system and listening to them.

3. From any phone using speech technologies (both recognition and synthesis) in an auditory mode. In this situation the users speak their requests for information, which are interpreted by the Mobile Assistant. Examples of requests are: “do I have any messages from John?”, “what’s on my calendar next Friday at 10,” or “play my voicemail messages.” The MA replies to their queries and reads them the requested messages or calendar entries.

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. Over 200 users at IBM Research are currently using this system to access messages and calendar information. The component of the system that was targeted by the study reported in this paper is the third component, which allows users to access messages in an auditory mode using speech technologies. While use of a SmartPhone allows users silent navigation of text-only (i.e., non-audio) messages, only a small percentage of the user population own these phones and access to the network is still sporadic.

3. METHODOLOGY

3.1. Experimental Design

We employed a within-subject design in order to maximize the power of the statistical tests. All participants experienced both the speech and the DTMF condition. In both conditions, the output modality (synthetic speech) and the tasks were the same. We created two test email accounts and calendars for the experiment, populating them with messages and appointments. The order of modality and test account used was counter balanced to eliminate any possible order effect. Each account had fifteen email messages. Sixteen participants (eight females and eight males) were recruited with a wide range in ages (late teens to over sixty). All participants except one, were naïve users with regard to speech systems. Tasks were representative of normal usage for the system (e.g., “what time is your meeting with David on Friday”).

3.2. Procedure

The participants took the study one at a time in a usability lab. They were told that the purpose of the study was to examine certain issues related to the use of a phone-based universal messaging and calendar system called the Mobile Assistant. Upon arrival in the lab, the participants were seated and given a booklet with the instructions on the first page. The instruction page described the purpose of the study and had a list of the required tasks. It also had the phone number to call to reach the system, the name of the test account they should use, and the password. For the DTMF condition, the instruction page listed all the keys and their corresponding functions.

The participants were told they would be interacting with the Mobile Assistant on the telephone to manage several email and calendar tasks. The experimenter showed the list of required tasks to the participant, made sure the participant understood what was involved, and showed him/her the data questionnaire that needed completing after each condition.

Participants were also given a sample interaction. On the instruction page for the DTMF condition, the sample was: “For example, to listen to the second email message in your inbox, press 5 and after listening to the message press the # key to hear the next message.” (See Table 1 for a complete listing of the DTMF functions that were shown to the participants).

Because the system accepts natural language input, it was not necessary (nor would it be feasible) to define everything that could be said to the system. However, in an effort to balance both conditions, the participants were given sample phrases to illustrate the types of things that would be understood. The sample phrases included on the task description were:

- How many messages did I receive yesterday?
- Do I have any messages from Peggy Jones?
- What’s on my calendar next Wednesday?
Table 1: DTMF keys used and their corresponding functions

<table>
<thead>
<tr>
<th>DTMF Key</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>* (star key)</td>
<td>Interrupt system output</td>
</tr>
<tr>
<td># (pound key)</td>
<td>Next message/Next day</td>
</tr>
<tr>
<td>0</td>
<td>Cancel current request</td>
</tr>
<tr>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Play phonemail</td>
</tr>
<tr>
<td>4</td>
<td>Play today's calendar</td>
</tr>
<tr>
<td>5</td>
<td>Play first email msg.</td>
</tr>
<tr>
<td>6</td>
<td>Delete a message</td>
</tr>
<tr>
<td>7</td>
<td>Repeat</td>
</tr>
<tr>
<td>8</td>
<td>Reply to a message</td>
</tr>
<tr>
<td>9</td>
<td>Forward a message</td>
</tr>
</tbody>
</table>

Participants were asked to complete each of the six tasks (once) in each condition. The list of tasks can be seen in Table 2. The only difference between the speech and the DTMF condition was the inbox that they were logging in to (and thus the messages contained in it) and the method of interaction used.

Both test inboxes were balanced for number of messages, type of message and total number of words for all messages. Also, each message was approximately the same length as the message in the same position in the other test inbox. Thus the first message in test inbox 1 had about the same number of words as the first message in test inbox 2, as did the second message, third message etc....

3.3. Measures

According to ETSI (the European Telecommunications Standards Institute), usability in telephony applications is defined as the level of effectiveness, efficiency and satisfaction with which a user achieves specific goals in a particular environment [4]. Effectiveness is defined here as how well a goal is achieved in a sense of absolute quality; efficiency as the amount of resources and effort that are used to achieve the specific goal; and satisfaction as the degree that users are satisfied with a specific system.

In this study, we examined all three elements of usability. We measured the effectiveness of a system by calculating a success rate for each user task. The amount of time to finish each task was used as a proxy measure for efficiency. In order to measure user satisfaction we had the users answer survey questions. Immediately after completing the six tasks in a condition, participants were asked to evaluate their interaction with the system by indicating how well certain adjectives described the interaction, on a scale of 1 to 10 (1 = “Describes Very Poorly”, 10 = “Describes Very Well”). Four adjectives—comfortable, exhausting (reverse coded), frustrating (reverse coded), and satisfying—were used to create an index we called “interaction satisfaction.” Five adjectives—boring (reverse coded), cool, entertaining, fun, and interesting—were used to create an “interaction entertainment” index. Finally, another four adjectives—artificial(reverse coded), natural, repetitive(reverse coded), and strained (reversed coded)—were used to form an “interaction naturalness” index.

4. FINDINGS

4.1. Efficiency of the system

We measured the efficiency of the system by looking at the success rate and the time for each task. The speech condition had lower success rates than the DTMF in 4 out 6 of the tasks, and took longer than DTMF in 4 out 6 also. (See Table 2 for average times in seconds for each task.) Part of the reason for the low success rates with the speech condition is that users tended to move on to the next task if they encountered substantial (more than 3) speech recognition errors while attempting a task. We believe that the longer times with speech were due to decoding latency, recognition errors and hesitation on the part of users who were not sure exactly what to say. The shorter times with speech for tasks 3 and 6 are due primarily to the direct access vs. sequential access dictated by the interaction style. For example, for task 3 the users could state something similar to “do I have any messages from” firstname lastname. With the DTMF interaction, the users had to listen to the header for each message and navigate to the next message before finding the message they were looking for. Similar arguments apply to the calendar task.

<table>
<thead>
<tr>
<th>Task</th>
<th>Speech</th>
<th>DTMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1: Log on</td>
<td>42 sec.</td>
<td>22 sec.</td>
</tr>
<tr>
<td>Task 2: Listen to the fourth message in the mail box</td>
<td>56 sec.</td>
<td>47 sec.</td>
</tr>
<tr>
<td>Task 3: Find out if they received a message from a particular person.</td>
<td>19 sec.</td>
<td>46 sec.</td>
</tr>
<tr>
<td>Task 4: Reply to the above message.</td>
<td>10 sec.</td>
<td>6 sec.</td>
</tr>
<tr>
<td>Task 5: Delete the message.</td>
<td>20 sec.</td>
<td>9 sec.</td>
</tr>
<tr>
<td>Task 6: Find out what time they are meeting with David Jones on Friday.</td>
<td>44 sec.</td>
<td>102 sec.</td>
</tr>
</tbody>
</table>

Table 2: Average times (in seconds) for task completion

4.2. User satisfaction

User satisfaction was evaluated through a series of survey questions, asked immediately after the use of each modality. For statistical analyses, we used a repeated measure ANOVA with modality as the repeated factor. There was no between-subjects factor.

Participants evaluated their interaction via speech modality ($M = 5.16; S.D. = 1.99$) as more satisfying than with DTMF ($M = 3.84; S.D. = 1.81$), $F(1, 15) = 4.35$, $p<.055$, more entertaining than with DTMF ($M = 3.69; S.D. = 1.47$), $F(1, 15) = 12.77$, $p<.01$, and more natural ($M = 3.65; S.D. = 1.63$), $F(1, 15) = 4.12$, $p<.06$ (see Figure 1 for a graphical representation of these results).

In response to the question “which of the two navigation methods did you prefer?” 69% of participants chose speech as their preferred modality. When asked why, participants
indicated for the most part that it is because using speech is easy and fun. The other main reason cited was that it frees up their hands and enables them to do multiple tasks at the same time. For those who preferred DTMF, the dominant reason was that it is less error-prone than speech.

Although the explanations proposed here are suggestive rather than conclusive, the findings underline the importance of examining speech interfaces from other perspectives in addition to efficiency maximization. Earlier work [7] comparing satisfaction and efficiency between a TTS-only system and one which mixed TTS and human speech, also found that the more efficient system (TTS-only) was not the most satisfying. While the current study reinforces how user satisfaction may not always correspond completely with efficiency measures, this may be a short-term effect. It is possible that given prolonged usage, the novelty factor of using speech could wear off and the preferred input modality might become DTMF. We are running a longer-term experiment with an internal deployment of the prototype. Since the time of the initial experiment, the end-to-end speech accuracy has increased to closer to 85%. Users currently have the option of using speech or DTMF for most functions. An examination of usage logs over time will show to what extent DTMF is used as an alternative to speech.

![Figure 1: Mean Values for Subjective Evaluation](image)

### 5. DISCUSSION

Our results indicate that users preferred spoken interaction to DTMF for the message retrieval system. This was in spite of lower efficiency (in 4 out of 6 tasks) in the speech condition and accuracy rates averaging only between 75% and 80%. This finding was very surprising to us given these low accuracy numbers (which were due to the early development stage of the prototype). Our prior experience with speech systems has shown that accuracy levels below 85% become highly unusable. Part of the explanation might be the “cool factor” of speech since most of the participants were novice speech users. In this case, we might expect to see abandonment of speech over time. Another explanation is certainly task-related since navigating through even a small number of messages in the sequential manner dictated by the use of DTMF, can quickly become tedious and repetitive. Lastly, there is the fun factor to consider. When interviewing a participant as to why he preferred the speech modality when his experience in the speech condition had been less than stellar he replied, “Well, I guess speech is just more fun.”

How can we design more successful speech interfaces? Clearly part of the answer is that speech systems need to have some element of fun, in addition to usefulness. Since speech recognition cannot currently boast 100% accuracy, nor will it be able to in the near future, it needs to compensate for this in other ways. The system used in this experiment is designed to be contrite and apologetic if it does not understand what the user is saying. It also makes (rather feeble) attempts at humor, is always polite, and takes the blame when accuracy problems strike. Since this factor was invariant across both conditions in the experiment, it is possible that the speech input modality was preferred simply because it seems more natural to speak to a system that is speaking to us. This finding would be in keeping with the media equation theory [9], which states that we react to new media (e.g., computer speech) with our “old brains”. Thus, we are conditioned to respond in kind when we hear speech, and the anthropomorphic characteristics of the system contributed to this effect.

### 6. REFERENCES


