ON THE DESIGN OF EFFECTIVE SPEECH-BASED INTERFACES FOR DESKTOP APPLICATIONS[1]

Jim Hugunin and Victor Zue
Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
hugunin@mit.edu — http://www.sls.lcs.mit.edu/~jjh

ABSTRACT

Is speech a useful input modality for applications where the user has easy access to a full-size keyboard and mouse? This study shows that a well-designed speech interface can be more effective than a standard desktop application's traditional interface. Subjects are able to build a set of three spreadsheet tables 50% faster using a spoken dialog interface, and they report significantly greater enjoyment in using that interface. However, these advantages cannot be achieved by simply bolting a speech recognition system onto an application's existing interface. We found that this latter approach led to an insignificant 4% increase in efficiency and a devastating 64% increase in errors compared to the standard keyboard and mouse interface. In short, speech-based interfaces have the potential to substantially improve our interactions with computers, but they require significant interface redesign to take advantage of the unique properties of speech.

1. INTRODUCTION

In recent years the research community has made tremendous progress towards producing accurate and reliable speech recognition systems. This progress suggests that widespread use of speech-based interfaces may soon become feasible. An outstanding question is what sorts of applications are these interfaces good for? Most people seem to agree that there are a number of obvious benefits to using speech in mobile situations to replace poor alternatives such as a touch-tone telephone keypad or a PDA's touch screen or tiny keyboard.

But what about traditional desktop applications? Can speech be a useful interface for applications where the user also has easy access to a full-size keyboard and mouse? Many studies [1,2] have shown that speech is ineffective as a direct replacement for the keyboard and mouse in existing applications. They attribute these failings to weaknesses in current speech recognition systems in terms of high error rates and/or large latencies. In this paper we suggest that these studies have failed to show a clear advantage for speech-based interfaces because their level of interface redesign has been insufficient.

We have performed a set of experiments demonstrating that in order to harness the power of speech it is not sufficient to bolt a speech interface onto an existing application. This will simply reveal the disadvantages of speech without exploiting its power. Introducing the spoken language features of dialogue and discourse into the application's interface is essential to harnessing speech's power as an input modality.

2. EXPERIMENTAL DESIGN

2.1. Three Ways to Build a Spreadsheet

The experimental study examines three different interfaces to a commonly used spreadsheet program (Microsoft's Excel 97). Each interface is used to build three different simple spreadsheet tables. These tables are taken from [1] with some slight modifications.

The baseline interface is the standard keyboard and mouse bindings that come with the application. This interface is the result of many years of development on spreadsheets in general and on Microsoft's Excel in particular. It should be a good example of what can be achieved with keyboard and mouse-based interfaces.

The second system bolts a speech interface onto the existing design of the application. It replaces the keyboard and mouse commands of Excel with spoken commands, but otherwise makes no changes to the application's design. The keyboard and mouse are inactive in this system, and the subject can interact with the program only by using spoken commands such as the following. "Two six five one" enters 2651 in the current cell; "select cell a two" selects the named cell in the table; and "sum row two through four" computes a function over a range of cells.

Finally, the third system redesigns the interface to the application to exploit the power and conventions of spoken language. The system conducts a dialog with the user about how to design their spreadsheet. It will prompt the user with likely tasks, such as, "Now enter the row labels for your table." It has enough knowledge of the table's structure to automatically select the appropriate cell for most entered values. It also has a level of intelligence that will let it guess and act on the user's intentions. When the user enters a row label as "Class Average", the system can recognize the word "average" as usually indicating a row which is computed based on other rows, and ask the user if they'd like to have this row computed by averag-

[1] This research was supported by DARPA under contract N66001-94-C-6040, monitored through Naval Command, Control and Ocean Surveillance Center.

ing down the table's columns. Finally, the system has the capability to support infinite undo's, allowing the user to easily correct the mistakes all too frequently caused by speech recognition errors.

2.2. System Implementation

For all three interfaces, the spreadsheets are produced using Excel 97. For the keyboard and mouse-based system, the user is interacting directly with Excel using its native interface. For the two speech-based systems, Excel is driven using it's OLE automation interface from a separate program which performs the speech recognition and understanding. All direct keyboard and mouse input to Excel is disabled under the speech-based systems.

Microsoft's Whisper [3] provides the core recognition engine for the two speech-based systems. The vocabularies in both interfaces are about 70 words, working in speaker-independent, continuous speech mode. The voice command grammar has an average utterance error rate of 11.6%. The voice dialog system takes advantage of constraints from the active dialog state to reduce this error rate to 8.4%. These error rates vary greatly across speakers from as low as 1% to as high as 23%. The system is running on a dual-processor 200 MHz Pentium Pro-based computer, which provides sufficient processing power to eliminate most latencies due to computation. This results in a half-second lag between the completion of an utterance and the recognized phrase.

A pilot study used on-screen text as the sole means of communicating with the user. We found that this failed to create the sense of an interactive dialog. In this study we use Microsoft's Agent [4] to provide an on-screen character for subjects to interact with. This agent (who speaks using a standard text-to-speech system) appears to greatly enhance the interactivity of the dialogs.

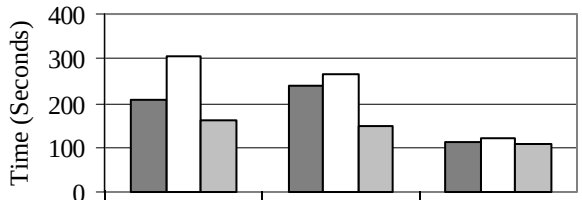
2.3. Subjects

Eighteen paid subjects were recruited at MIT for an experiment in speech recognition. All subjects were explicitly required to be native speakers of American English; *not* to be computer scientists; and *not* to be expert touch typists. Each subject used all three interfaces in the experiment. The order of the interfaces was permuted to ensure that all interface orderings were used equally over the experimental pool of eighteen subjects.

The subjects were given a post-experiment questionnaire to assess their experience with spreadsheets, typing, and voice input. According to these self-evaluations (and in agreement with observed behavior) all but one subject was an average or very good typist; their experience with spreadsheets ranged from first time users to experts; and their experience with voice input ranged from first time use to some experience with voice-mail systems.

2.4. Training

Subjects were trained in the use of each interface in the context of solving two practice problems designed to exercise all of the techniques needed to solve the three "real" problems in the experiment. A sheet of example commands was provided for each interface to suggest to the subjects the sorts of things they might want to type, mouse, or speak to produce the spreadsheets. During these practice problems, the experimenter was listening to the subject from a separate room, and watching the subject's responses on a slave screen. The subject was allowed to ask any questions she might have in order to complete the practice problem. In addition, the experimenter would offer suggestions when the subject seemed excessively confused.



	Keyboard and Mouse		Voice Commands		Voice Dialog	
	μ	σ^2	μ	σ^2	μ	σ^2
Problem 1	207	153	236	83	113	33
Problem 2	306	192	264	78	119	34
Problem 3	160	83	146	44	106	59
Average	225	141	215	59	113	31

Figure 1: Time to complete each problem as a function of interface.

This design allowed subjects to discover their own personal styles for dealing with each interface, while at the same time making sure that each subject learned enough to complete the rest of the experiment. The amount of help needed from the experimenter varied greatly depending on the subject and the interface. Virtually all subjects needed some help to complete the voice command practice problems; all of the inexperienced spreadsheet users needed help with the keyboard and mouse interface; unsurprisingly, the least help was needed for the voice dialog-based interface.

3. RESULTS

3.1. Time to Complete Task

Figure 1 shows the average time for each subject to complete each spreadsheet problem using each interface. This only counts time actually spent working on building the spreadsheet by each subject. Subjects were given as much time as they wanted to familiarize themselves with each problem before beginning work.

As expected, bolting voice commands onto the existing interface does not lead to any significant increase in efficiency. Only a 4% reduction in time to complete the three

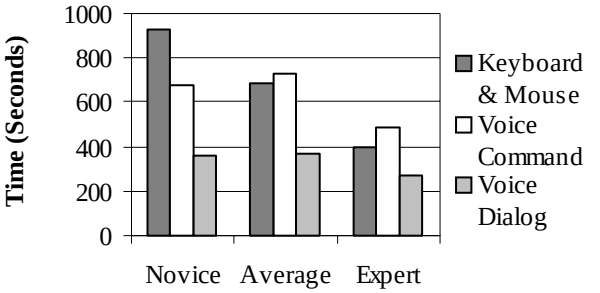


Figure 2: Time to complete all problems as a function of spreadsheet expertise and interface.

problems is observed for the voice command interface. Using a 1-tailed paired T-Test [5] this difference is not statistically significant (p=0.37).

In contrast, the spoken dialog interface is 50% faster than the keyboard and mouse. This increase in performance is statistically significant (p=0.0005). Although this speed increase varies dramatically across subjects, from 7% to 74%, it is interesting that no subject is faster using the keyboard and mouse than the voice dialogs.

3.2. Errors in Completed Tables

The subjects in this experiment left a number of errors in their completed spreadsheet tables. These errors are different from the large number of speech recognition or typing errors that occurred during the building of the tables. These are the errors that the subjects failed to notice in their tables before submitting them as complete. There were 93 possible errors for each interface, and subjects had an average error rate of around 1%. This rate is higher than one would expect from actual users completing important work, and partially reflects the lack of adequate motivation for subjects to produce error-free tables.

Subjects made an average of 0.78 final errors across the three problems when using the standard keyboard and mouse interface. They made slightly more using voice dialogs (0.89) but this difference is not significant (p=0.41). Subjects made 64% more errors using voice commands than using the keyboard and mouse (1.28). This difference is weakly significant (p=0.12).

3.3. Effects of Spreadsheet Expertise

The subjects' expertise with spreadsheets was assessed through a post-experiment questionnaire asking them to rate their experience from 1 (first-time) to 10 (expert). These responses were grouped into novice (1-2), average (3-6) and expert (7-10) users. Figure 2 shows how the time to complete all three problems varies across interface and spreadsheet experience.

Several clear trends emerge from this data. As expected, in the keyboard and mouse condition the time to complete the spreadsheet tables decreases dramatically with the subjects' spreadsheet experience. With the voice command interface, the time to complete the tables also

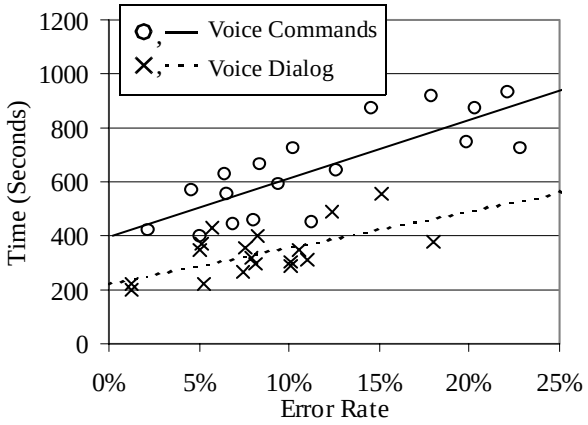


Figure 3: Time to complete problems as a function of speech recognition error rate. Points are plotted for each subject, lines give least squares fit.

appears to decrease with increased experience, but this result is much weaker. Particularly interesting is that while the voice command interface appears to be clearly more efficient than keyboard and mouse for novice spreadsheet users, it loses this advantage for experts.

The voice dialog users show virtually no dependence of efficiency on spreadsheet experience. This is to be expected from the complete redesign of the interface for this system. Despite the dramatic improvements in subjects' performance with the keyboard and mouse interface with experience, even the expert spreadsheet users found the voice dialogs provided a significant advantage over the keyboard and mouse system.

The results for novice users merit a little explanation. These results include the single computer illiterate subject in this experiment. This was an older gentleman who had only tens of hours experience working on a computer (which is still far greater experience than any subject had with voice input). It took him more than 30 minutes to complete the three spreadsheet problems with the keyboard and mouse. This is 2.4 times as long as the next slowest subject. Removing his data from the results will reduce the degree of difference between novice and more experienced users, but won't change the above conclusions. On the other hand, if this experiment had sought out more such subjects to truly explore the novice end of the spectrum, the magnitude of these differences would be greatly increased.

3.4. Effects of Recognition Accuracy

Not surprisingly, the speed with which subjects can complete problems using the speech-based interfaces is directly related to the underlying recognition error rate. This relationship is shown in Figure 3 for both the voice command and voice dialog interfaces. Clearly these systems work optimally only when the recognition error rate approaches 0%. Improving the recognition accuracy of any speech-based system is one of the most obvious ways to improve its usefulness and enhance its power.

3.5. User Enjoyment

Finally, and perhaps most importantly, users reported enjoying their experience significantly more with the voice dialogs than with either the keyboard and mouse or the voice commands ($p=0.0007$). Only three out of eighteen subjects reported that the keyboard and mouse system was more enjoyable than the voice dialogs. There was no significant difference in reported enjoyment between the voice command and the keyboard and mouse systems ($p=0.43$).

4. DISCUSSION

4.1. Limitations of the Current System

Despite the encouraging performance and user satisfaction results reported in this paper, the voice dialog system is clearly a demonstration of the concept, not a potential competitor to any real spreadsheet applications. In order to achieve reasonable performance with Whisper [3], the system uses a 70 word vocabulary and a severely constrained grammar. This means, for example, that the grammar believes the year has only five months. It believes that there are only 8 possible first names and 8 possible last names in the world. It can't conceive of any number larger than 9999. Perhaps most importantly, it only understands (and therefore can create) the very simplest class of spreadsheet tables.

4.2. Recruiting Subjects

Recruiting an appropriate set of subjects is a challenge for experiments comparing novel interface techniques to standard systems. All but one subject in this experiment had hundreds or thousands of hours experience using a keyboard and mouse to communicate with a computer. Several of the subjects in this experiment had hundreds of hours experience specifically using Excel's standard keyboard and mouse-based interface. In contrast, the most experience any subject had with speech-based input was less than one hour, and most subjects had no experience with talking to computers whatsoever.

It will remain virtually impossible to recruit subjects with substantial experience using speech recognition systems until these systems become powerful enough to be widely used in the general population. What remains a challenge is designing that next generation of systems without having such an experienced subject pool.

4.3. Recognition Errors

One of the unfortunate properties of a speech recognition system is its stochastic nature. For any given utterance, there is a finite chance that the computer will make a mistake in interpreting what the user said. This is very different from the errors made while typing, which are no less frequent, but which are clearly the user's fault.

One of the situations where these errors cause the most trouble is with very naive users, who are still trying to learn how to use an unfamiliar system. These users have a strong tendency to assume that a recognition error means that they failed to state a command properly. They can therefore become very confused in the process of simultaneously learning what command to issue, and how to issue it so the computer will understand.

Another danger of speech recognition errors is their potential to become catastrophic. This is the phenomenon where no matter what the user tries, she is incapable of getting the computer to successfully recognize a given phrase. In our system, a set of N-best recognitions were re-sorted based on previous errors to avoid this problem, but clearly a more robust solution is needed for a production system.

5. SUMMARY AND FUTURE WORK

This experiment suggests that spoken language technology has the potential to create revolutionary new interfaces that can simultaneously increase both the user's enjoyment and efficiency for interacting with their computer. It also shows that these improvements are not going to come from merely bolting a speech recognition system onto an application's existing keyboard and mouse-based interface. Instead, researchers must learn how to design a new generation of user interfaces that take the unique power and constraints of spoken language into account.

The voice dialog system used in this experiment is an example of a particular system designed to take advantage of this new input modality. At this point we can make no claims as to how well these techniques will generalize to other applications or even to more sophisticated spreadsheets. Further work is clearly needed to discover a general set of design principles for speech-based interfaces that can be applied to the rich variety of computer applications. Failure to do so will only encourage designers to fall into the comfortable trap of applying existing GUI design techniques to a fundamentally different input modality.

6. REFERENCES

[1] R. Damper and S. Wood, "Speech versus keying in command and control applications", *International Journal of Human-Computer Studies*, Vol. 42, pp. 289-305, 1995.

[2] K. Molnar and M. Kletke, "The impacts on user performance and satisfaction of a voice-based front-end interface for a standard software tool", *International Journal of Human-Computer Studies*, Vol. 45, pp. 287-303. 1996.

[3] X. Huang et al., *"Microsoft Windows Highly Intelligent Speech Recognizer: Whisper"*, IEEE International Conference on Acoustics, Speech, and Signal Processing, 1995.

[4] Microsoft Agent, <http://www.microsoft.com/workshop/prog/agent/>

[5] B. Winer, "Statistical Principles in Experimental Design", McGraw-Hill Book Company, 1962.