A Comparison of Voice Controlled and Mouse Controlled Web Browsing

Kevin Christian* kevin@cs.umd.edu

Bill Kules*+ wmk@cs.umd.edu

Ben Shneiderman*^ ben@cs.umd.edu

Adel Youssef* adel@cs.umd.edu

*Department of Computer Science
University of Maryland at College Park
College Park, MD 20742

^Human-Computer Interaction Laboratory
Institute for Advanced Computer Studies
Institute for Systems Research
University of Maryland at College Park
College Park, MD 20742

+Takoma Software, Inc.
7006 Poplar Avenue
Takoma Park, MD 20912

ABSTRACT

Voice controlled web browsers allow users to navigate by speaking the text of a link or an associated number instead of clicking with a mouse. One such browser is Conversa, by Conversational Computing. This within subjects study with 18 subjects compared voice browsing with traditional mouse-based browsing. It attempted to identify which of three common hypertext forms (linear slide show, grid/tiled map, and hierarchical menu) are well suited to voice navigation, and whether voice navigation is helped by numbering links. The study shows that voice control adds approximately 50% to the performance time for certain types of tasks. Subjective satisfaction measures indicate that for voice browsing, textual links are preferable to numbered links.

Keywords

Human-computer interaction, user interfaces, voice browsers, voice recognition, web browsing

INTRODUCTION

Information contained on the World Wide Web is inaccessible to many people. The web is primarily a visual medium that requires a keyboard and mouse to navigate, and this disenfranchises several types of users. People who lack motor skills to use a keyboard and mouse find navigation troublesome. Visually impaired users can not read the display. People who do not have access to an Internet-capable computer have difficulty even accessing the World Wide Web, and those who temporarily cannot use a traditional web browser (for example, because their eyes or hands are occupied or because they are not near their computer) are at a minimum inconvenienced.

Speech recognition and generation technologies offer a potential solution to these problems by augmenting the capabilities of a web browser. A voice browser is a web browser with at least one of the following capabilities:

- Can render web pages in an audio format (speech generation)
- Can interpret spoken input for navigation (speech recognition)

A number of voice browsers are on the market, and more are under development. Conversational Computing's Conversa is a web browser that accepts speech input, but renders the pages in the traditional visual manner [18]. The Home Page Reader, from IBM, renders web pages in audio format, but accepts commands only via the keyboard's number pad [20]. PipeBeach is a system that affords both audio rendering of web pages as well as speech input. LIASON, from Siemen's, Inc., is a system designed for use while driving an automobile[25]. Systems specifically designed to accommodate telephone-based browsing include Lucent's PhoneBrowser, Siemen's DICE, and 1-800-Hypertext [1,4,25]. Other systems are application-specific. VADAR, from BBN, allows users to track shipments over the world wide web, while Talk'n'Travel, also from BBN, is an interface for commercial-travel websites that allows users to access flight and train schedules [20]. The GALAXY project at MIT is a system that will access the web to find information in response to a user's queries [21].

Users with temporary or permanent motor impairments stand to gain much from such products. A web browser that can render web pages in audio format will be of obvious use to the blind, and navigating by voice obviates the need for keyboard and mouse navigation. Additionally, people whose eyes and hands are otherwise engaged may still be able to conveniently access the web. For example, someone will be able to get directions via the web while driving their car.

Voice browsers open up new possibilities for bringing the content of the web to a larger segment of the population. A voice browser potentially makes the telephone capable of Internet access. Since the number of households with telephones is far greater than number of households with internet-capable computers, it stands to reason that the number of people with internet access will increase greatly once voice browsers become widely available. Moreover, telephones are far less expensive than computers, so voice browsing will help open up the World Wide Web to low-income users. Also, sales of wireless telephones are flourishing; voice browsing may enable the owners of such phones to browse the web wirelessly from virtually any location.

There are several challenges facing designers of voice browsers. First, a web page rendered with voice output is inherently a temporal medium. In a visually presented web page, many different images, tables and the like can be presented on the screen at the same time, in a spatial format that is quickly and effectively processed by the human perceptual system. Spoken text, however, can only be presented one word at a time. While some research has gone into using multiple, simultaneous, non-speech sounds, reading of screen contents can only occur in a sequential, linear fashion.

Second, formulating speech commands and processing speech output consumes the users' short-term and working memory and conflicts with tasks such as planning and problem solving that depend on the same forms of memory. Visual information is processed in a separate system, permitting parallel operations. A study by Karl et al. noted that subjects had more difficulty memorizing symbols when commands to manipulate those symbols were issued by voice than when commands were issued via the keyboard or mouse [11].

Third, there is the inevitable recognition error involved in speech recognition systems. Recognition error refers to situation in which the user speaks one word but the system chooses another as the best match. After nearly 30 years of research in the area of natural language recognition, the best systems remain relatively unsophisticated. A recent system boasts a recognition rate of 93% with a vocabulary of 1000 words, and even this requires background lexical and syntactic knowledge [8]. While users tend to view recognition error as a sign of immature technology, some researchers believe that recognition error is inevitable [3].

Research has not proven the effectiveness of speech recognition as a general-purpose input mode. A study by Van Buskirk and LaLomia had subjects complete tasks involving navigation in a graphical user interface (GUI). In half the tasks subjects used spoken commands and in the other half subjects used a keyboard and mouse. They found that voice navigation took approximately twice as long as traditional navigation [23]. Earlier studies produced inconsistent and conflicting results [6,12,24].

Speech input can be useful in certain situations. Research into multimodal interfaces indicates an distinct advantage to using speech as an input mechanism. The aforementioned study by Karl et al. showed that using speech to issue commands to a word processing application, while using the keyboard for text entry and the mouse for direct manipulation, significantly sped up task time. Similarly, Mignot et al. showed that the addition of spoken commands to direct manipulation (via a touchpad) greatly reduced the task performance times of their subjects [13].

Two common threads run through both of these experiments. First, in both tasks, the number of commands that could be issued via spoken input is relatively small. Second, the users spoke very short sentences. For example, in the Karl study, speech input was only used for commands such as "File Open", or "Save". Even the Van Buskirk and LaLomia study, which demonstrated a significant performance decrease associated with voice navigation, noted that, "the best tasks for speech input were tasks in which the user has to issue brief commands using a small vocabulary".

There is some theoretical and experimental justification for this. A study by Poock [17], cited in the Karl paper, demonstrated a clear advantage for issuing commands by voice over issuing commands via keyboard. Oviatt, in [15], showed that both the length of spoken commands and lack of structure in input format is proportional to the number of disfluencies made by the user. A speech disfluency is any type of unnatural disruption in normal speech, such as a repetition, filled pause (e.g. "umm"), self-correction, or false start. Oviatt claims that long sentences lead to more complicated plans for formulating input and that these more complicated plans are more prone to errors. Also, if the input grammar is unstructured, users have more options in formulating their input, which leads to more disfluencies.

Oviatt's focus on the types of errors and the ease with which they can be corrected reflects the current trend in speech input research. This research suggests that error detection and correction is the crucial factor in determining task completion times. Danis and Karat found that when using speech-recognition systems, the types of errors users commit are fundamentally different from errors committed with other input styles [3]. This tends to confuse users who are not accustomed to recognizing and correcting such errors. A study by Karat, et al., noted that, "when subjects made errors in keyboard-mouse text entry, they tended to correct the error within a few words of having made it. In contrast, some subjects made specific mention of not being as aware of when a misrecognition had occurred and needing to 'go back to' a proofreading stage for the speech tasks" [10]. This same study noted that subjects made almost four times as many errors using speech recognition for transcription tasks as they did when using keyboard-and-mouse. A follow-up study [7] investigated user strategies for correcting errors. They identify two common strategies: spiral depths, where users re-dictate misinterpreted words, and cascades, where misrecognition (frequently of commands) caused addition errors, which needed to be corrected before the original error could be dealt with. Similar effects are noted in [16].

The characteristics of web navigation may be advantageous to voice-controlled navigation. A small number of commands can provide the navigation functionality common to visual browsers, and current technology is effective for small vocabularies. Web navigation commands are typically short, such as "go back", "follow link", "refresh", and "read next frame." Short commands such as these have a very high degree of structure; in fact, there is almost no grammar to speak of, as each command maps to exactly one combination of sounds. (Some voice browsers, notably IBM's Home Page Reader and telephone browsers, ignore this problem altogether by using the number pad for input.)

There are potential limitations, though, that may reduce the utility of voice browsing. Although there is a small set of commands, users must typically speak the text of the link (i.e., the text that a mouse user would click) to follow the link. A web page author can use virtually any string of characters to represent links, which creates a potentially unlimited universe of valid voice commands with very little structure, not all of which are valid English words. These must be spelled out letter by letter for the speech recognition system to properly recognize the link. The error rates associated with such large, complicated, unstructured command set may be quite large.

The central question we sought to answer is whether navigating the World Wide Web by voice is a viable alternative to traditional mouse-based navigation. Would it produce results similar to those found in the Van Buskirk and LaLomia study (slower) or those found by Karl et al. (faster)? Based on the literature and our experience, we hypothesized that speech navigation will be noticeably slower than navigating with the mouse, but not quite twice as slow.

Our experiment also attempted to discern when numbered links are more helpful than text links as navigational aids. We hypothesized that, due to the simplicity of the spoken commands when using the numbered links, navigation with voice and numbered links would be faster and less error prone than navigation with voice and text links. Finally, we anticipated that users would appreciate the voice control capability because of the flexibility and novelty, and that this would be reflected in higher subjective satisfaction ratings for the voice methods.

We limited our experiment to three common web navigation patterns: the hierarchical menu, the linear slide show, and a two dimensional panning map (no zooming). We used a single voice browser product, Conversa, which renders pages visually and supports voice as well as the traditional "point and click" technique with a mouse. The user typically traverses links by speaking the text of the hyperlink (i.e., the text that a traditional user could click with a mouse). Image maps, links containing text which is not English, and links in densely packed regions of many links are assigned a number (sequentially in a top-to-bottom, left-to-right manner), and the user speaks that number to follow the link.

Conversa does not require or support user-specific speech recognition training. It provides a limited set of preferences to customize the tool. For speech recognition, the user can adjust for the speaker's voice pitch (male, female or child) and speech recognition precision (from lenient to strict in 5 increments). It is positioned as a mass-market product for use by both experts and novices.

EXPERIMENT

Hypotheses

We hypothesized that navigating the web by voice introduces a noticeable delay in completion times of tasks, but that the time to complete tasks via voice browsing would be at most twice that of traditional mouse-based browsing. Furthermore, using numbered links would be faster and less error-prone than using textual links. With regard to subjective satisfaction, we hypothesized that users will prefer voice-based browsing to mouse-based browsing.

The independent variable was the style of navigation used by the subjects. We had three dependent variables: task completion time, error rate, and subjective satisfaction. The task completion time is defined as the time taken to complete a given task. The error rate is defined as the number of times a subject has to repeat a command due to an error on the part of the voice recognition software. A subject's subjective satisfaction was measured by a questionnaire given to the subject after he or she completed the tasks.

This was a 1x3 experiment. The first treatment was mouse-only navigation. For this treatment, subjects navigated the web in the traditional manner. The second treatment was voice navigation with text links. Subjects followed links by speaking the hypertext. The third treatment was voice navigation with numbered links; Conversa numbered each link on a given page, and navigation was accomplished by speaking the number. The keyboard was not used in any of the treatments

Subjects

A total of 18 subjects were used; 12 of these were male, and 6 female. Slightly over half of these subjects were affiliated in some way with the computer science department at the University of Maryland. One was a faculty member, three of these were instructors, five were graduate students, and one was a staff member. The others were acquaintances of one of the experimenters. Eleven subjects were between the ages of 10 and 29, five were between the ages of 30 and 39, and two were between the ages of 50 and 59. All had significant experience using computers and web browsers, but none had any experience with voice browsers. All subjects spoke English without a noticeable accent.

Materials

The web browser we used was Conversa, produced by Conversational Computing (http://www.conversa.com).

Conversa is a full-featured browser, supporting both voice and mouse navigation. Also, Conversa automatically numbers links that are represented by images. Subjects used a Labtec C-324 microphone to provide voice input.

Web pages were specially constructed for this experiment. There were three tasks for each treatment, each designed to evoke a particular pattern of navigation. The same set of pages were used for mouse and voice navigation with textual links. The start pages for each task were accessed through a common home page.

The first set of tasks used a 4x4 tiled map. A large map was split into 16 equal sized pieces, and tasks involved moving the "frame of focus" around the landscape. Users moved the frame of focus by indicating to the browser that was to go either north, south, east, or west. Figure 1 shows a web page from this set that used text links. A sample task was "Starting from Detroit, following the red line, what is the name of the destination city located at the end of the line?"

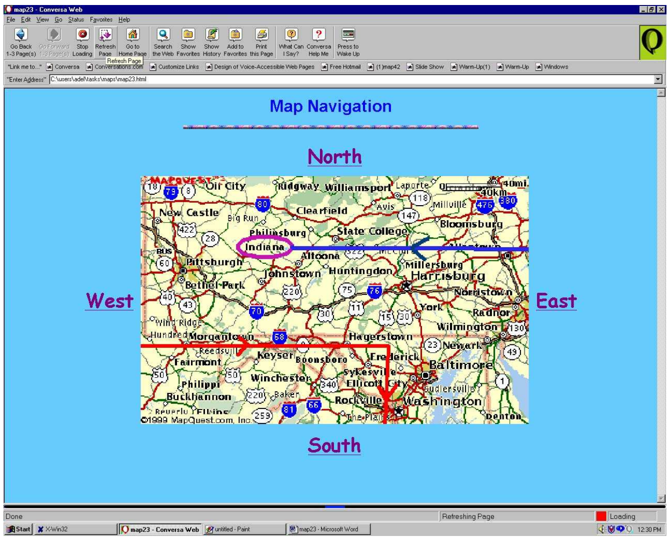

Figure 1. Sample map

The second set of tasks was a slide show. Ten web slides were created; each displayed a random number. Figure 2 shows a slide with numbered links (the actual numbers appear inside of the small yellow balloons). Tasks involved navigating through the slides and relaying to the administrator the number on the target slide. A sample task for this set of pages was "Go to the last slide, and then go back four pages. What is the number on the sixth slide?"

The third set of tasks was a hierarchical-tree style menu. Zaphiris and Mtei studied the differences in task completion times between short, fat trees and tall, narrow trees. They constructed 64 web pages, each of which contained information about the nation of Cyprus. Using these 64 pages as leaves, they constructed trees of varying heights and branching factors. We chose the 4x3 set of pages so that our tree would be equally poised between depth and breadth. Tasks involved looking for information about Cyprus. Subjects were asked a question about Cyprus, and then beginning at the root page, they navigated through the tree to locate the leaf page containing the requested information. Sample tasks for this set of pages were "In 1992, who was Cyprus' Minister of Finance?", and "What was Cyprus' national product in 1992". Figure 3 shows a menu with numbered hyperlinks, shown in yellow balloons.

Figure 2. Sample slide

We performed a pilot study and made minor revisions to the materials and procedures based on the results. We found that users would start the tasks before they had finished reading the instructions, so we changed the instruction to specifically direct them to read all questions before starting. We also simplified the tasks slightly and reworded several questions that were found to be confusing.

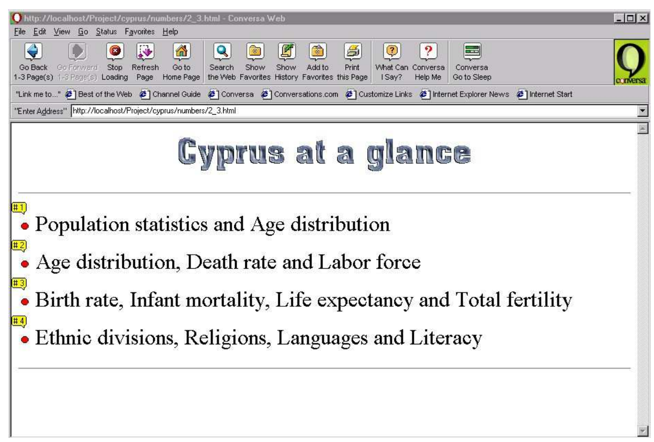

Figure 3. Sample page from hierarchy showing numbered hyperlinks

Procedures

Most of the test procedure was managed using paper checklists and forms. This avoided requiring users to interact with a test harness while also performing the tasks, although it required one test observer for each test. Users were asked to complete the subjective satisfaction questionnaire on-line after the test.

Prior to each test, the sequence of the three treatments was selected and the checklist was prepared. All six permutations of treatment sequences were used (three subjects per sequence) to compensate for order effects.

Subjects were initially welcomed and given a brief description of voice browsing. We then described the tasks and asked them to sign the consent form. Detailed instructions in the use of the voice browser software were provided, along with a review of the icons used for typical mouse operation (e.g. Back, Home Page). Users were asked to perform the sample tasks on a set of warm up pages, and then given as much time as desired to continue familiarizing themselves with voice-browsing techniques.

When the users indicated they were ready, the main part of the test began. The users were asked to read the two questions associated with the first treatment and task, then indicate when they were ready. The test observer would then tell the user to begin, and would start the timer. While the user performed the tasks, the observer counted errors. When the user returned to the experiment home page, the observer stopped the timer. This was done for all nine tasks.

Problems

During the pilot study, we noticed that users would begin a task before completely reading and understanding the questions. To avoid including the users' reading times in the task times, we specifically asked users to read and understand the questions before beginning the task. Users did not always do this, and even when they did, they sometimes re-read the questions immediately after starting a task. This certainly contributed to some variation in task times.

When performing the slide task for the first time, users often misused the "Go Back" command when trying to navigate to a preceding slide. They should have used the "Previous" command instead. The linear layout of the slides contributed to this problem because of the dual meanings of forward and back in this context. After realizing the difference, the users did not make this mistake in the second and third treatments of the slide task.

There were several specific problems with the maps that caused difficulty for subjects. The map quality was marginal, especially the text. One subject commented that he missed a landmark (Detroit) because of the text graininess. Users were also distracted by changes in the alignment (or registration) of the map segments as they panned. They occasionally went back to the previous segment to check their progress.

RESULTS

The experimenters analyzed the task completion times, error rates, and subjective satisfaction of each participant using Microsoft Excel.

Completion Time

A single factor analysis of variance (ANOVA) was performed on the task completion times for each treatment. The results show that navigating slide shows with a mouse is faster than navigating with voice commands. This result was significant at alpha = .05, $f(2,51) = 4.93$, $p = .011$. Task completion times for navigating hierarchical menus with a mouse were also faster than when navigating with voice commands. These results were statistically significant at alpha = .05, $f(2,51) = 12.82$, $p < .001$). However, when navigating tiled maps, the results were not statistically significant at alpha = .05, $f(2,51) = .27$, $p = .76$).

For the tasks with statistically significant results (slide show and hierarchical menus), a set of three paired t-tests were performed. The results show significant differences between mouse and both types of voice browsing, but no significant differences between the two voice treatments.

Table 1 shows mean completion times with the standard deviation in parentheses.

	Mouse	Voice/ Text	Voice/ Numbers
Slide show	35.1 (24.9)	52.1 (14.5)	56.6 (23.9)
Map	54.0 (30.1)	52.6 (18.1)	58.1 (19.6)
Menu	39.8 (13.8)	58.9 (18.3)	68.3 (19.0)

Table 1. Mean task completion times in seconds with standard deviation in parentheses (n = 6 per cell)

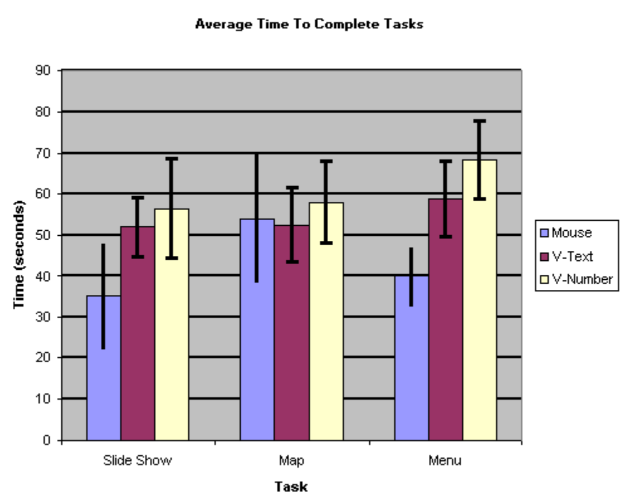

Figure 4. Mean task completion times

Error Rates

Mouse errors were excluded from the error rate analysis. A paired t-test was performed on the two voice treatments. With respect to the number of missed commands, the results were not statistically significant at alpha = .05.

With respect to the number of misinterpreted commands, the results obtained showed that the number of misinterpreted commands is negligible. Because of the low error rates, no further statistical analysis was warranted.

Subjective Satisfaction

A single factor ANOVA of the subjective satisfaction questionnaire shows a preference for text links over numbered links. Results for all three questions were statistically significant at alpha = .05, $f(2,51) = 15.97$, $p < .001$, $f(2,51) = 20.34$, $p < .001$, and $f(2,51) = 15.78$, $p < .001$, respectively. A set of three paired t-tests were performed. In terms of overall reaction to browsing style and navigating to the desired page, the results show that the voice-text treatment has a statistically significantly higher rating than voice-numbers treatment. However, in terms of tool ease-of-use, there was no statistically significant difference between the two treatments.

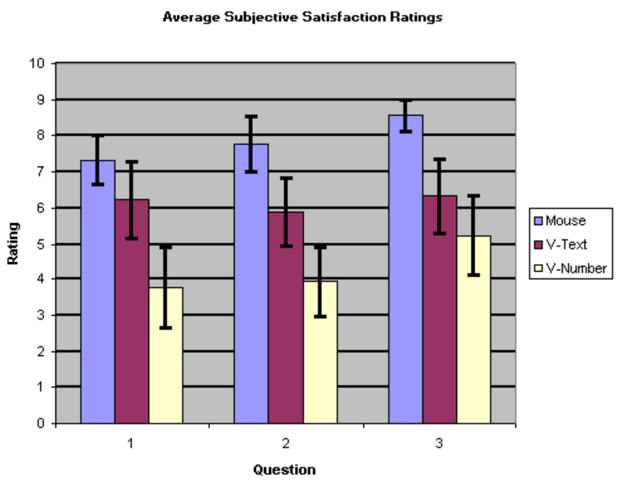

Figure 5. Subjective Satisfaction Ratings

Question 1 asked users for their overall reaction to the browsing style. Question 2 asked users how easy they found navigation with a particular tool. Question 3 asked users how easy they found it to use the tool. All questions were graded on a scale of 1 to 9, where in question 1, a 1 signified "frustrating" and 9 signified "satisfying", and in questions 2 and 3, a 1 indicated "difficult" and 9 indicated "easy".

DISCUSSION

Subjective Satisfaction

There was a significant difference between subjective ratings for the text and number-based voice browsing styles. This was corroborated by user observations during the test and their comments afterwards. We observed users executing an extra cognitive step when using the numbers. They had to determine the numeric value of the balloon number associated with the desired link before they could speak the number and activate the link. We noticed pauses and "double-takes" as users mapped the text to a number, then spoke the number. For the map tasks, where the links were known *a priori*, the user could simply speak the command (e.g., North) as soon as they decided what direction to move, without needing to read the link. When speaking the text of a link, users could simply speak their choice without needing to make the separate conversion of "text to numbers". After the test, users commented on the difficulty of using the numbers.

The quantitative results to questions one (overall reaction to the browsing style) and two (ease/difficulty of navigating to the desired page) do not support the hypothesis that users would prefer the voice alternatives, although many users commented positively on the technique. This may be a result of a lack of familiarity with the voice browsing technique combined with the wording of the questions. More specifically worded questions might have allowed us to better quantify the positive comments that we heard. The advantage for question three (learning to use the tool) is understandable, since all users were familiar with using the mouse for browsing and none of the users had used a voice browser before.

Performance

There were significant differences in task performance times by treatment for two of the tasks, the slide show and the Cyprus data. For these tasks, the voice browsing technique took on average 1.5 times as long as the mouse technique. This is consistent with our hypothesis.

As noted above, the balloon numbers added an extra cognitive step, which may have contributed to the time difference between the text and numbered voice treatments, although these differences were not statistically significant.

The average times for the map tasks were not significantly different. It is possible that the treatments are equally effective for the map navigation task. We observed that users had difficulty counting landmarks while navigating, suggesting that the cognitive demands of counting and navigating were in conflict, as noted in [11], causing more user errors and extending performance times. It is also possible, however, that the results were confounded by several factors, including user confusion over the questions, poor text quality and ragged alignment of map segments.

Errors

There were no statistically significant differences in error rates. Error rates for misinterpreted commands were low. It appears that even when the speech recognition engine is set near the lenient end of the scale (the default configuration), the software is conservative and is more likely to reject a possibly correct match than to make an incorrect match.

Expert Users

We became familiar with Conversa during the course of the experiment, and we measured the performance of one author in the performance of the tasks to suggest what expert performance might be like. No speech recognition errors (missed commands or misinterpreted commands) occurred. Overall, task times were less than for other subjects, presumably because of the author’s experience. Otherwise, the results correlate with the rest of the study. For this user, the slide and hierarchical menu tasks took about one-third longer when using voice control, while the map task times were about the same for all treatments.

CONCLUSION

The results of this experiment suggest that motor-impaired users who speak English without an accent will be able to use voice control to navigate the World Wide Web. They will not need to train the speech recognition software to their specific voice. They may initially experience some voice recognition errors when speaking links, but with a modest amount of experience those errors should become rare.

Impact for Practitioners

When creating web pages for voice navigation, designers should ensure that hyperlinks are easily spoken English text. Similarly, image links should be used sparingly. Non-English links and images are converted to numbered links, and the results show that numbers are harder for subjects to learn and use and extended task completion times.

When numbered links are unavoidable and appear on several consecutive pages (for example, navigation bars that are used throughout a web site), ensure that they appear in the same relative ordering, so that the numbers are consistent between pages.

Speech recognition for input is less precise than the mouse, and links that sound similar could inadvertently be activated when using voice control. Designers should therefor word links on a page so that they are short (a few words) and aurally distinct.

Developers of voice browsers should consider alternatives to numbered links. Rather than using numbers to activate image links, voice browsers could display the text of the ALT attribute in an image link and accept that instead of or in addition to a number.

Suggestions for future researchers

One obvious direction for future research is to explore more common website architectures. This experiment looked at slide shows, tiled maps, and hierarchical menus. Other common types include index pages (a page consisting entirely of a large number of links), and zoomable images (Yahoo! maps are an example). Also, this study simply attempted to identify types of web pages that might be better suited to textual links and which might be better suited to numbered links. Better insight into why certain types of web pages appear to favor one style over the other would be helpful.

Other aspects of web browsing are problematic for voice control. Entering a specific URL (i.e., one that is not linked to by the current page) is difficult and error prone. Each individual letter of the URL must be spoken using the military codes (alpha, bravo, charlie, etc.). Even though the browser displays a table of letter codes, the process is unwieldy at best. The browser also displays a list of recently linked to URLs that match the partial URL being entered, and these links are numbered using the balloon technique. This helps alleviate some problems when entering URLs, but there is still no good mechanism for entering a new URL. The mechanisms for filling in forms are similarly awkward. More convenient ways to manipulate checkboxes, radio buttons, and drop-down menus should also be investigated.

Using speech to control the browser necessarily limits other conversation. This is potentially surmountable by distinguishing between commands directed to the browser from speech that is part of other conversation. Much like participants in a telephone conversation recognize changes in tone and volume to detect (and ignore) side conversations that the other party may be having, speech recognition software could be configured to similarly respond only to, for example, a lowered tone of voice. This would permit a user to carry on normal conversation without inadvertently activating a link. Non-verbal cues (as mentioned in [2]) could also be used to infer when commands are being directed to the browser.

Refinements to the Theory

Voice control adds approximately 50% to the performance times for simple navigation tasks that are focused on rapid navigation through multiple links. Tasks that require less frequent navigation and those in which the links are known in advance (e.g. map navigation) should experience little time difference.

Voice commands do introduce cognitive overhead. After users identify which link they want to follow, formulating and enunciating the correct voice command takes longer than moving a mouse to the desired location and clicking. This overhead seems to be slightly more severe when numbered links are used as opposed to text links. When a user spots an appropriate link, it is easier to simply read rather than associating a number that has no inherent relevance to the context of the link.

ACKNOWLEDGEMENTS

We would like to thank Anne Rose for facilitating our use of the HCI Laboratory. We would also like to thank our pilot subjects, who provided helpful comments and suggestions. Lastly, our sincere thanks to the subjects who graciously agreed to take part in this study.

REFERENCES

1. Brown, Michael K., Stephen C. Glinski, Bernard P. Goldman, and Brian C. Schmult. "PhoneBrowser: A Web-Content-Programmable Speech Processing Platform." Position Paper for The W3C Workshop on Voice Browsers, Cambridge, MA (October 1998), http://www.w3.org/Voice/1998/Workshop/Michael-Brown.html
2. Cassell, J. "Embodied conversational interface agents"; Communications of the ACM 43, 4 (Apr. 2000), 70-78
3. Danis, C. and Karat, J. "Technology-Driven Design of Speech Information Systems", Symposium on Designing Interactive Systems. ACM: New York (1995), 17-24.

4. Glashan, Scott Mc. "Standards for voice browsing." Position Paper for The W3C Workshop on Voice Browsers, Cambridge, MA (October 1998), http://www.w3.org/Voice/1998/Workshop/ScottMcGlashan.html

5. Goose, Stuart; Wynblatt, Michael; and Mollenhauer, Hans. "1-800-Hypertext: Browsing Hypertext With a Telephone", Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia, Pittsburgh, PA (June 1998).

6. Haller, R.; Mutschler, H.; and Voss, M. "Comparison of Input Devices for Correction of Errors in Office Systems." Proceedings of INTERACT '84, London (1984).

7. Halverson, C. A.; Horn, D.; Karat, C.; and Karat, J. "The Beauty of Errors: Patterns of Error Correction in Desktop Speech Systems", Proceedings of INTERACT '99 (1999), 133-140.

8. Hwang, M-Y; Hon, H-W; and Lee, K-F. "Modelling Between-Word Coarticulation in Continuous Speech". Proceedings of EUROSPEECH '89, Paris (1989), 5-8.

9. James, F. "Lessons Developing Audio HTML Interfaces, Assets." Proceedings of the Conference on Assistive Technologies, Marina Del Ray, USA (April 1998).

10. Karat, C.; Halverson, C.; Horn, D.; and Karat, J. "Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems", Proceedings of ACM CHI99: Human Factors in Computing Systems (1999), 568-575.

11. Karl, Lewis; Pettey, Michael; and Shneiderman, Ben. "Speech Activated versus Mouse-Activated Commands for Word Processing Applications: An Empirical Evaluation." International Journal of Man-Machine Studies, vol. 39, 4 (Oct. 1993), 667-687.

12. Martin, G. L. "The Utility of Speech Input in User-Computer Interfaces". Journal of Man-Machine Studies vol. 30, 4 (1989), 355-375.

13. Mignot, Christophe; Valot, Claude; and Carbonell, Noelle. "An Experimental Study of Future 'Natural' Multimodal Human-Computer Interaction". Proceedings of ACM INTERCHI'93 Conference on Human Factors in Computing Systems Proceedings (1993), 67-68

14. Motorola, VoxML: The Mark-up Language for Voice [online]. http://www.voxml.com.

15. Oviatt, S. "Interface Techniques for Minimizing Disfluent Input to Spoken Language Systems". Proceedings of ACM CHI'94 Conference on Human Factors in Computing Systems, v.1 (1994), 205-210.

16. Oviatt, S. and Van Gent, R. "Error Resolution During Multimodal Human-Computer Interaction", Proceedings of the Fourth International Conference on Spoken Language Processing, University of Delaware and AI DuPont Institute, New York, 204-207.

17. Poock, G. K. "Voice Recognition Boosts Command Terminal Throughput". Speech Technology, I, 2, 36-39.

18. Robin, Michael and Hemphill, Charles. "Considerations in Producing a Commercial Voice Browser." Position Paper for The W3C Workshop on Voice Browsers, Cambridge, MA (October 1998).

19. Robin, Michael and Jim Larson. "Voice Browsers: An introduction and glossary for the requirement drafts." [online] W3C Working Draft, (December 1999). http://www.w3.org/TR/voice-intro/

20. Stallard, David. "BBN Position Paper on Conversational Web Access." Position Paper for The W3C Workshop on Voice Browsers, Cambridge, MA (October 1998). http://www.w3.org/Voice/1998/Workshop/DaveStallard.html

21. Strom, Nikko. "Position Paper for the W3C Workshop: Voice Browsers". Position Paper for The W3C Workshop on Voice Browsers, Cambridge, MA (October 1998). http://www.w3.org/Voice/1998/Workshop/MIT-SLS.html

22. Thatcher, Jim; Jenkins, Phil; and Laws, Cathy. "IBM Special Needs Self Voicing Browser". Position Paper for The W3C Workshop on Voice Browsers, Cambridge, MA (October 1998). http://www.w3.org/Voice/1998/Workshop/PhilJenkins.html

23. Van Buskirk, R., and LaLomia, Mary. "A Comparison of Speech and Mouse/Keyboard GUI Navigation". Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, v.2 (1995), 96.

24. Visick, D; Johnson, P; and Long, J. "The Use of Simple Speech Recognizers in Industrial Applications". Proceedings of INTERACT '84, London (1984).

25. Wynblatt, Michael and Goose, Stuart. "Towards Improving Audio Web Browsing". Position Paper for The W3C Workshop on Voice Browsers, Cambridge, MA (October 1998). http://www.w3.org/Voice/1998/Workshop/Siemens.html