Design Decisions for a Voice Navigation System

MARIA MILENKOVIC
IBM Corporation, 1555 Palm Beach Lakes Blvd., W. Palm Beach, FL 33401
maro_m@vnet.ibm.com

ALAN J. HAPP
IBM Corporation, P.O. Box 12195, 3039 Cornwallis, Research Triangle Park, NC 27709
alanhapp@vnet.ibm.com

JAMES R. LEWIS
IBM Corporation, 1555 Palm Beach Lakes Blvd., W. Palm Beach, FL 33401
jimlewis@vnet.ibm.com

Abstract. Voice navigation systems allow users to interact with computer applications by voice. In this paper we present human factors research that evaluated design alternatives for major user interface components. In the first study we evaluated the appropriateness and discriminability of four sets of icons that were candidates for functions in the **Voice Toolbar** of the navigation system. In the second study we studied two different ways of organizing and presenting voice commands in the **What Can I Say** window. The key finding was that the alternative organization seemed to improve task performance. The results provided the basis for design recommendations.

Introduction

Significant advances in accuracy and performance in automatic speech recognition have resulted in several commercial voice navigation and dictation systems. A voice navigation system allows a user to interact with the desktop and other computer applications by voice. In this paper we present human factors research that evaluated design alternatives for major user interface components and guided some of the implementation decisions for a voice navigation system. Together with dictation, the voice navigation system is part of IBM's family of speech products. This is a speaker independent, continuous speech, large vocabulary, command and control system. A user does not need to "train" the system to recognize his or her voice and can use a large number of continuous phrases to control computer applications. The design reflects the need to converge three lines of speech offerings that were based on different speech technologies, developed by different groups within IBM, and intended for different operating environments.

Several assumptions underlie the design of the voice navigation system. First, we assumed that voice navigation is part of a multimodal interface, where speech is a valid input modality along with keyboard, mouse, or pen. Although it attempts to capitalize on the advantages of speech, the design is not for a "speech only" interface. Voice navigation and dictation represent two modules of functionality that need to complement each

other with seamless integration. We assumed two levels of users: an "average" and a more "experienced" user. The "average" user is familiar with business applications needed for a job. A more "experienced" user is familiar with prevalent operating environments and conventions (GUIs, desktop metaphors) and wishes to use voice to complement or augment typical tasks.

Design Objectives

The design objectives for the voice navigation system included the following:

- Specifying the visual elements of the interface.
- Specifying the user interaction with the system.
- Designing a voice toolbar to integrate the modular speech functions of navigation and dictation at the level of the interface.
- Supporting a visually consistent and unified "look and feel" for this family of speech products across platforms and operating environments.

The final design of the voice navigation system reflects the influence of several factors:

- Previous expert evaluation of competitive voice navigation systems. The evaluation yielded a set of desirable design elements and usability criteria for a "best of breed" navigation system.
- User interface guidelines applied to the design of user interfaces in general and those supporting speech as an input modality.
- Human factors research to evaluate and validate design alternatives.

Two main user interface components received particular consideration during the design process: the **Voice Toolbar** and the **What Can I Say** Window.

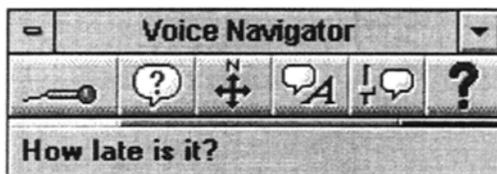The **Voice Toolbar**, shown in Fig. 1, consists of several user-configurable elements:

Figure 1. An illustration of the **Voice Toolbar**.

- A row of buttons that provides immediate access to speech functions, such as "What Can I Say", for a list of valid voice commands, "Where Can I Go", for a list of voice-accessible programs, "Begin Dictation" to start the dictation function, and "Voice Settings", to access voice recognition properties and settings.
- A speech volume indicator.
- A speech feedback area, where the system echoes what was recognized or misrecognized .

The **What Can I Say** window displays a list of valid voice commands the user can say at any time. The organization of the displayed information is of utmost importance, as users consult the **What Can I Say** window not only to review valid voice commands, but to see how to complete a task using voice.

Human Factors Studies

Throughout the design and implementation process, human factors research helped evaluate design alternatives and guide decisions. In this section we will report the findings from two studies. The first study evaluated four sets of icons which were candidates for functions in the **Voice Toolbar** (Ehrlich et al., 1995). The second study (Milenkovic et al., 1995), evaluated the usability of an alternative organization of the **What Can I Say** window.

Study One: Evaluating the Appropriateness of Four Sets of Icons for a Speech Recognition Toolbar

In the first study, we assessed the appropriateness and discriminability of four sets of icons. The *Microphone Off* icon indicates that the microphone is off and the voice recognition system can not respond until the microphone is on. The *Microphone Asleep* icon indicates that the voice recognition system does not respond until the user says "wake up". The *"Where Can I Go"* icon opens a window with a list of programs and applications the user can access by voice. The *Dictation mode* icon is a visual indicator that the application is in dictation mode rather than navigation or command and control mode.

The IBM Icon Reference Book (1991) provided guidelines for testing the appropriateness of icons. The testing method used a paired comparison procedure to assess both appropriateness and discriminability of the icons. The dependent measure was a score based on

the subject's preference for one or the other icons or both (i.e., there was no difference in preference).

Method

Subjects. Fourteen subjects participated in the study. One group consisted of seven people hired from a local agency, described as experienced application users, but not experienced with speech products. The second group consisted of IBM employees who were current users of voice navigation and dictation products.

Icons. Table 1 shows the four sets of icons that the participants evaluated. The *Microphone Off* and *Microphone Asleep* icons differ in the orientation of the microphone component (diagonal or horizontal) and the background color used (either gray or yellow). Two of the icons with the horizontal microphone were yellow and one had the letters "Zzz" in black shown above the microphone (icon #3 and icon #4). The *Dictation Mode* icons all used the voice bubble and colored characters: red in icon #2, blue in icon #3. In the *Where Can I Go* icon set, the arrows were red. In icon #2 of the set, the question mark and the window title were teal[1].

An Authorware™ program presented the icons. The program randomly selected each set without replacement and randomly presented pairs of icons without replacement. Each icon pair appeared in a separately labeled display. For a set of 4 icons, e.g., the *Microphone Off* set, there were 6 different displays. The top of each display showed instructions and the concept that the participant was to evaluate. The display showed the pair of icons in a numbered list.

Procedure. Participants completed a short demographic questionnaire that collected information on computer use and, specifically, use of speech recog-

nition products. When ready, a participant pressed the space bar to start the paired comparison test. For each comparison, a participant stated whether the icon labeled "1" or "2" better represented the concept being evaluated. A participant could respond "both" or "neither" if there was no preference for either of the icons (i.e., they were equivalent). The experimenter recorded all responses from a participant on a response sheet. Participants worked at their own pace. They completed the comparisons for a set of icons before starting a new set; after each set, the experimenter asked the participant for suggestions of other alternatives.

Dependent Measure. The test methodology prescribes the following algorithm for scoring a participant's responses: "for each icon pair displayed, the most appropriate icon received a score of "2" and the least appropriate icon received a score of "−1". When neither icon was preferred over the other, each icon in the pair received a score of "1". (IBM Icon Reference Book, 1991, pp. 15–21). For the icon sets with four icons, the range of possible scores was 9, from "−3" to a maximum of "+6". For the Dictation Mode icon set, the minimum score was "−2" while the maximum score was "+4". For the "Where Can I Go" icon set, the maximum preference score was "+10", while the lowest possible preference score was "−5".

Results

Microphone Off Icon Set. Table 2 shows the mean and standard deviation of the preference scores of the *Microphone Off* icon set for both groups of participants. A repeated measures analysis of variance (ANOVA) with experience and icon as factors showed a marginally significant main effect of icon ($F(3, 36) = 2.28$, $p = 0.10$) and experience by icon: ($F(3, 36) = 2.36$, $p = 0.09$). Overall, participants

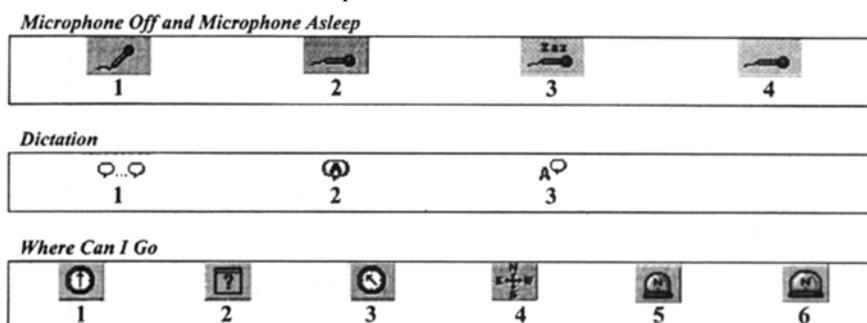*Table 1.* Four icon sets used in the experiment.

Microphone Off and Microphone Asleep

Dictation

Where Can I Go

Table 2.   Preference scores for the icons evaluated for the concepts Microphone Off and Microphone Asleep.

		Microphone Off		Microphone Asleep	
Icons	*Experience Level*	*Mean*	*Std Dev.*	*Mean*	*Std Dev.*
A	HI	1.29	3.40	-1.14	3.48
	lo	-0.86	3.39	-1.00	2.00
B	HI	4.43	2.22	0.86	1.21
	lo	2.14	2.16	0.57	0.53
C	HI	-0.14	3.08	4.29	3.40
	lo	3.00	3.46	5.86	0.38
D	HI	1.71	1.98	2.57	1.90
	lo	3.14	2.67	2.29	1.50

preferred Icon B for *Microphone Off*. The interaction was marginally significant because the pattern of preference differed for experienced and less experienced users. Experienced users strongly preferred Icon B, but less experienced users expressed satisfaction with Icons B, C, and D.

Microphone Asleep Icon Set.   Table 2 also shows the mean and standard deviation of the preference scores of the *Microphone Asleep* icon set for both groups of participants. A repeated measures analysis of variance (ANOVA) with experience and icon as factors showed a significant main effect only for the icon factor $(F(3, 36) = 16.57, p < 0.001)$. Overall, participants preferred Icon C to represent *Microphone Asleep*.

Dictation Mode Icon Set.   Table 3 shows the mean and standard deviation of the preference scores of the Dictation Mode icon set for both groups of participants. A

Table 3.   Preference scores for the icons evaluated for the concept dictation mode.

Dictation

Icons	*Experience Level*	*Mean*	*Std Dev.*
A	HI	-1.71	0.76
	lo	1.86	1.57
B	HI	2.57	1.51
	lo	1.29	2.21
C	HI	2.86	1.35
	lo	0.29	2.14

repeated measures analysis of variance (ANOVA) with experience and icon as factors showed a significant effect for the icon factor, and a significant interaction $(F(2, 24) = 3.55, p = 0.04,$ and $F(2, 24) = 9.58, p < 0.001,$ respectively). Overall, participants preferred Icon B. An examination of the mean preference scores shows that the significant interaction occurred because the icons most preferred by the experts (B and C), were least preferred by the novices (who most preferred A).

Where Can I Go Icon Set.   Table 4 shows the mean and standard deviation of the preference scores of the *Where Can I Go* icon set for both groups of participants. A repeated measures analysis of variance

Table 4.   Preference scores for the icons evaluated for the concept Where Can I Go.

Where Can I Go

Icons	*Experience Level*	*Mean*	*Std Dev.*
A	HI	1.86	1.86
	lo	3.71	2.21
B	HI	2.00	4.24
	lo	4.86	1.68
C	HI	2.57	1.99
	lo	1.71	3.30
D	HI	9.86	0.38
	lo	8.14	2.16
E	HI	3.00	1.73
	lo	2.57	4.24
F	HI	3.86	2.34
	lo	1.71	2.87

(ANOVA) with experience and icon as factors showed a significant main effect only for the icon factor ($F(5, 60) = 12.34$, $p < 0.001$). Participants strongly preferred Icon D.

Discussion. One goal of this study was to facilitate the adoption of an appropriate icon for selected functions of a speech product. Four sets of icons were evaluated: *Microphone Off, Microphone Asleep, Dictation,* and *Where Can I Go.* The decision to recommend one of the icons over the other rests not only on the results of the statistical analyses but also on design tradeoffs that affect the look and feel of the entire interface. One tradeoff was that the responses of the experienced speech recognition users were given more importance than those of the less experienced users. There were two reasons for this. First, the experienced users have a fuller model of speech recognition tasks so selecting icons that they prefer might foster that model in novice users, resulting in faster learning. Second, novice users are more likely to have the help "bubbles" turned on so that the system presents the function of the icon in text to the user. As novices learn the system, they may turn that help off, and at that point, they should have the most appropriate icon available. Another tradeoff is that any individual icon must fit with the look and feel of the interface. This means that even with a strong preference by users for an icon, the design team may make changes in the look of the icon to be sure it is consistent with the other icons in the application or system. Research has shown that the concept represented is more important than details such as color (Miller et al., 1992) or prototypical features (Byrne, 1991). The results of the study, along with other tradeoffs, suggested the following icons be used in the Voice Tool bar:

Microphone Off   *Microphone Asleep*

From an end user and system perspective, the *Microphone Off* and the *Microphone Asleep* icons must be considered together. The *Microphone Off* icon has a gray background while the *Microphone Asleep* icon has a yellow background. The novice users liked the icons with the yellow background better than the icons with the gray background regardless of the concept being evaluated. However, it would be better to take advantage of a difference in color of the icons. The experienced speech recognition users clearly preferred the recommended icons which have different backgrounds. The result of the ANOVA for the *Microphone Off* concept suggested trends for significant differences among the icons and differences depending on the type of user (the significant interaction). The recommended *Microphone Off* icon reflects the decision to use the icon that was slightly less preferred by the novice users since the system could provide help tags (text that pops up with the icon's label) and the recommended icon would help the novices learn the preferences of the experienced users.

Dictation

Interestingly, this icon was the most preferred by the experienced speech users but the least preferred by the novices. Two reasons support the recommendation to use this icon. First, it is most similar to the graphic on the button that opens the IBM Voice Type dictation window. Users should benefit from the consistency. Secondly, by using the icon preferred by the experienced users we might help teach the novice users to see the system more like an experienced user. Help tags, on initially by default, teach users what the functions are while they are learning the system.

Where Can I Go

Both groups of users endorsed the concept of a compass crosshair icon for the *Where Can I Go* button. (None of the participants mentioned it, but the directions for E and W are, in fact, mislabeled. The actual icon will have the compass directions properly labeled!)

In summary, the icon test methodology worked well for the evaluation of these icons. For small sets of icons this procedure can be done quickly and without fatiguing the participant. However, when the number (n) of icons increases, the sheer number of paired comparisons, which increases as $(n * (n - 1))/2$, makes the test a chore for the experimenter and the participant. Often, the real question is whether users find the set of all icons in a product appropriate and discriminable. When this is the case, methods used by Lewis (1993a), Fullerton and Happ (1993), or Lin (1992) should be considered.

*Second Study: Human Factors Evaluation
of an Alternative "What Can I Say"
Window Organization*

One current use for continuous speech systems is to enable people to issue verbal commands to control computer applications. This allows the use of the natural communication mode of speech. However, systems that lack perfect grammatical parsing and interpretation of the speech input (in other words, all modern commercial computer speech systems), require users to know what phrases the speech-enabled application can accept as valid commands. Thus, most command-and-control continuous speech systems provide, on command, a list of the phrases that the system can currently understand—a **What Can I Say** window. When a user isn't sure what to say next, he or she can browse the list of commands in this window.

The objective of the study was to evaluate the usability of an alternative organization of the **What Can I Say** window for the IBM™ Continuous Speech System™ (ICSS™), as displayed in the Human-Center, the multimodal user interface prototype for

IBM's Personal Computer Power Series™. Specifically, we sought to determine whether the alternative organization improves a user's ability to find a desired command. In both organizations, voice commands are organized and displayed under headings that indicate logical groupings. A small button with a "+" sign in front of the heading indicates that the group can be expanded or reduced to the principal heading by clicking the plus sign or by saying the name of the heading. For hiding the contents of the group, the user can click the button now displaying the minus sign, or say the name of the heading again. Previous research (Chimera and Shneiderman, 1994) indicates that a comparable "expand/contract" scheme or a multipane interface scheme, vs. a fully expanded interface, produces significantly faster times for information retrieval tasks. Relative to the initially designed organization, the alternative organization has fewer high-level categories, with the categories organized to focus more on likely user tasks. Figure 2 illustrates the appearance of the originally designed **What Can I Say** window, and Fig. 3 shows the alternative organization.
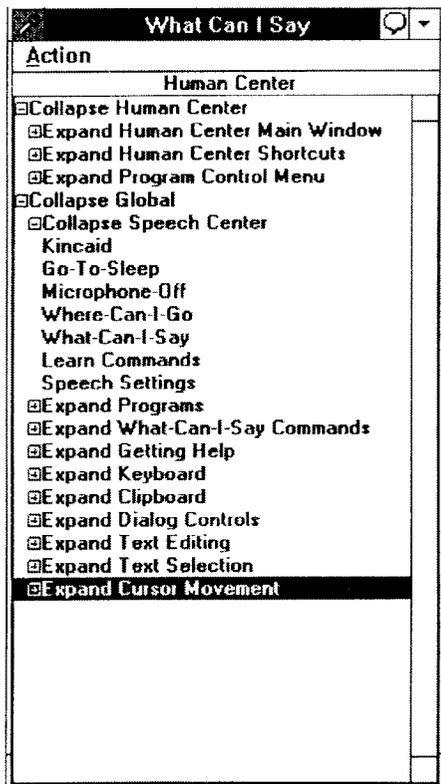
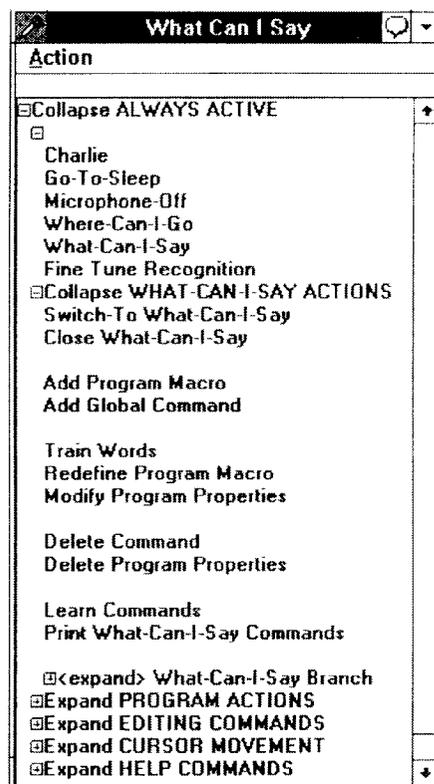*Figure 2.* An illustration of the initially designed **What Can I Say** window.

Figure 3. An illustration of the alternative **What Can I Say** window.

Method

Subjects.   Eight IBM employees (4 male and 4 female) participated in the study. All participants were native American English speakers and familiar with graphical user interfaces and IBM's voice recognition products.

Equipment.   The Human Center ran on an IBM Personal Computer Power Series Model 6015 (with the 601/66 MHz processor) using Windows NT[TM] as the operating system.

Procedure.   There were two types of tasks used in this study[2]. For the Explicit voice command tasks, the participants received a card which contained a voice command for which they were to search. In the Scenario tasks, participants read a short scenario describing a navigation task. They then searched for the appropriate voice command needed to complete the task. Each participant completed five of each type of task (Explicit, Scenario) for each type of **What Can I Say** organization (Current, Alternative). A Greco-Latin design (Lewis, 1993b) was used to counterbalance the order of presentation of task sets (a, b), type of task, and type of **What Can I Say** window. Table 5 shows the explicit phrases and scenarios used in this study.

To begin the experimental session, the experimenter explained the procedure to the participant. For each task: The experimenter demonstrated how to complete the task and handed the participant a card describing the task. The participant then searched for the appropriate voice command in the **What Can I Say** window. The experimenter recorded the time required to complete the search (Task Time), the number of times the participant clicked the mouse to expand or contract a high-level heading (Heading Clicks), and the number of times the participant clicked the mouse to move the information in the window (Scrolling Clicks).

The participants completed the tasks in sequence for each **What Can I Say** window organization as assigned by the Greco-Latin experimental design. After completing the tasks for one **What Can I Say** organization, there was a short break as the experimenter loaded the other **What Can I Say** organization. After participants finished all the tasks, they completed a short questionnaire which allowed them to comment on their experience.

Table 5.   Descriptions of the experimental tasks.

Number	Explicit voice commands	Scenario
1a	Move Window Right 20	You want to move to the end of the line that you're typing on. Find the command for this.
2a	Always on Top	Someone has come into your office and you want to switch off your microphone so the navigator won't pick up your conversation. Find the command for this.
3a	Back Tab 20	You want to activate the Escape key by voice. Find the command you need to use.
4a	Previous Program	You want to activate the Tab key 10 times. Find the command that allows you to do this.
5a	Top of Document	You are tired of looking at the What Can I Say window. Find the command that allows you to close it.
1b	Enter-Key	You are typing out a filename to save it and you want to find the command that changes from insert to typeover mode.
2b	Next Field 20	You are reading through a paper you've written in Write[TM]. You want to check the references at the end. Find the command that moves you to the very end.
3b	Select Complete Document	You want to delete the line you're typing, but first you need to select it. Find the command that allows you to select it.
4b	Speech Center Help	You are creating a large drawing document and you want to move down a page. Find the command that lets you move down one screen.
5b	Page-Up 20	You are constantly opening the same two applications together and you want to create a voice command that will do this for you. Find the command that allows you to do this.

Results

Search Time.   Figure 4 shows the results for the time required to search for the target phrase as a function of type of **What Can I Say** organization (Current, Alternative) and Stimulus Condition (Explicit, Scenario).   For the Explicit task, participants found the target phrase, on average, 36% faster with the alternative organization. For the Scenario task, the reduction in search time was an average of 23%. Although these effects appear relatively large, they were not statistically significant. The effects of stimulus condition and the organization by stimulus interaction were also non-significant. The reasons for the variability in search times are the small sample size and a strong, significant learning effect across trials, especially from trial 1 to 2. Part of what we have learned from conducting this study is that, given a small sample, it would be useful to provide more practice than we did to stabilize performance before starting the experiment proper.

Mouse Clicks.   The number of mouse clicks on headings within the **What Can I Say** window is another measure of how easy or difficult it is to find a specific voice command within the **What Can I Say** window. When clicking on a heading, the heading expands and reveals the valid voice commands grouped underneath. The assumption is that a greater number of mouse clicks
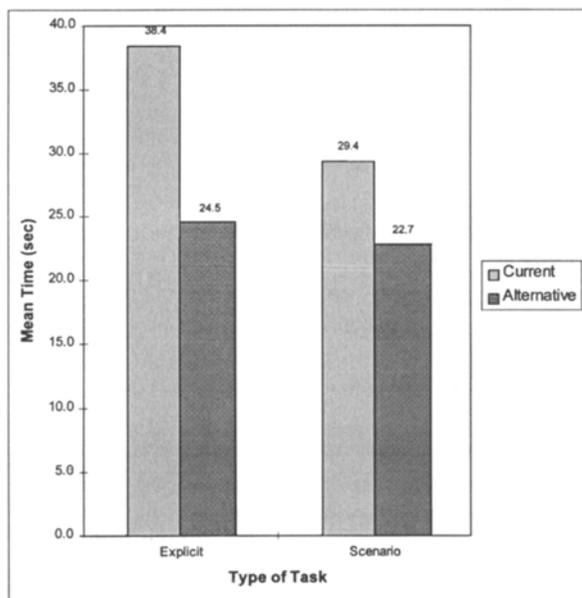
indicates a greater difficulty in locating a command because the participant has expanded several headings to find a particular command. The number of scrolling clicks was a similar measure.   It reflects how many times a participant clicked to scroll the whole **What Can I Say** window while trying to locate a command. The key finding is that the total number of mouse clicks was equal between the two organizations Thus, it is not reasonable to attribute differences in search times to differences in mouse click requirements between the two organizations.

Commands That Were Difficult to Locate.   To identify which commands were most difficult to locate in the alternative organization, we performed item analyses to identify:

- Commands with higher than average search time
- Commands with higher than average heading clicks
- Commands with higher than average scrolling clicks

These criteria indicated the potential for improvement in the location of several commands. For example, five commands (*Top of Document*, *Microphone Off*, *Escape Key*, *Speech Center Help*, *Select Line*) had measurements below average for all three criteria. Thus, these commands would be prime candidates for relocation in a subsequent redesign.

Discussion.   The main finding of this study was that the alternative organization seemed to improve task performance, with an average reduction in search time of 36% for the first type of task and an average reduction of 23% for the second type of task. Thus, the results suggest that the alternative organization is better than the initially designed organization. In addition to supporting the adoption of the alternative **What Can I Say** window organization, the data point to specific ways to further improve the alternative organization by relocating certain commands (identified in the Results section).   Failure to reach statistical significance for differences in search time is probably due to the small sample size and high performance variability. Because the apparent effect size was large, we did not replicate the study to increase the sample size in an attempt to achieve statistical significance.

Figure 4.   Mean search times as a function of the **What Can I Say** organization and type of task.

Conclusions

Traditional human factors methods and processes were applied while iterating through the design of a voice

navigation system and two of its major user interface components, the **Voice Toolbar** and the **What Can I Say** window. The evaluation of four sets of icons for discriminability and overall appropriateness provided the basis for recommending one icon in each set for the speech functions under consideration (*Microphone Off, Microphone Asleep, Dictation mode, and Where Can I Go*). The alternative **What Can I Say** window organization, with fewer higher level categories and categories organized to focus more on likely user tasks, seemed to improve task performance. We plan to pursue further exploration of alternative ways to structure and present valid voice commands to the user. Overall, subsequent usability enhancements to the organization of the **What Can I Say** window, as well as to the logical flow and sequencing of the more difficult tasks (creating a voice macro), will be incorporated in subsequent design iterations of the voice navigation system.

Acknowledgments

We wish to express our appreciation to Jennifer Ehrlich and Lise Schneider who helped to design the experiment and collected the data for evaluating the appropriateness of four sets of icons. We also wish to express our appreciation to Ron Van Buskirk, who participated in the design of the alternative **What Can I Say** organization, helped to design the experiment, and collected the data for the second study.

Trademarks

IBM, Personal Computer Power Series, IBM Continuous Speech System, and ICSS are trademarks of the International Business Machines Corporation. Authorware is a trademark of Macromedia Inc. Windows NT and Write are trademarks of Microsoft Corporation.

Notes

1. The icons were designed for the US product. National Language Support issues should be taken into consideration for international products.
2. The tasks were not necessarily directly comparable, but the design of the experiment and subsequent analyses did not require them to be comparable. Instead, they were designed to cover a reasonable range of tasks.

References

Byrne, M.D. (1991). The misunderstood picture: A study of icon recognition. In *Proceedings of Human Factors in Computing Systems (CHI '91)*. New York: Association for Computing Machinery, p. 493.

Chimera, R. and Shneiderman, B. (1994). An exploratory evaluation of three interfaces for browsing large hierarchical tables of contents. *ACM Transactions of Information Systems (USA)*, 12(4):383–406.

Ehrich, J., Happ, A.J., and Schneider, L. (1995). Evaluating the appropriateness of four sets of icons for a speech recognition tool (Tech. Report TR54.897). Boca Raton, FL: IBM.

Fullerton, S. and Happ, A.J. (1993). A user-oriented test of icons in an educational software product. In G. Salvendy and M.J. Smith (Eds.), *Proceedings of the Fifth International Conference on Human-Computer Interaction (HCI International '93)*. Amsterdam: Elsevier, pp. 44–49.

IBM Corporation (1991). *IBM Icon Reference Book*. (SC34-4348-00). Cary, NC: Author.

Lewis, J.R. (1993a). An icon usability evaluation procedure with application to personal communicator icons (Tech. Report TR54.792). Boca Raton, FL: IBM.

Lewis, J.R. (1993b). Pairs of Latin squares that produce diagram-balanced Greco-Latin designs: A BASIC program. *Behavior Research Methods, Instruments and Computers*, 25:414–415.

Lin, R. (1992). An application of the semantic differential to icon design. In *Proceedings of the Human Factors Society 36th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society, pp. 336–340.

Milenkovic, M., Lewis, J.R., and Happ, A.J. (1995). Human factors evaluation of an alternative "*What Can I Say*" window organization for ICSS (Tech. Report 54.907). Boca Raton, FL: IBM.

Miller, M.E., LaLomia, M.J., and Happ, A.J. (1992). Icon search processes: Effect of icon color and category cue on eye movements, identification time, and accuracy (Tech. Report TR54.659). Boca Raton, FL: IBM.