Speech-Activated versus Mouse-Activated Commands for Word Processing Applications: An Empirical Evaluation

Lewis Karl (COMSAT Laboratories), Michael Pettey (University of Maryland) and Ben Shneiderman (University of Maryland)

July, 1992

Send correspondence to:

Lewis R. Karl
COMSAT Laboratories
23500 COMSAT Drive
Clarksburg, MD 20871

Despite advances in speech technology, human factors research since the late 1970s has provided only weak evidence that automatic speech recognition devices are superior to conventional input devices such as keyboards and mice. However, recent studies indicate that there may be advantages to providing an additional input channel based on speech input to supplement the more common input modes. Recently the authors conducted an experiment to demonstrate the advantages of using speech-activated commands over mouse-activated commands for word processing applications when, in both cases, the keyboard is used for text entry and the mouse for direct manipulation. Sixteen experimental subjects, all professionals and all but one novice users of speech input, performed four simple word processing tasks using both input groups in this counterbalanced experiment. Performance times for all tasks were significantly faster when using speech to activate commands as opposed to using the mouse. On average, the reduction in task time due to using speech was 18.67%. The error rates due to subject mistakes were roughly the same for both input groups, and recognition errors, averaged over all the tasks, occurred for 6.25% of the speech-activated commands. Subjects made significantly more memorization errors when using speech as compared with the mouse for command activation. Overall, the subjects reacted positively to using speech input and preferred it over the mouse for command activation, however, they also voiced concerns about recognition accuracy, the interference of background noise, inadequate feedback and slow response time. The authors believe that the results of the experiment provide guidance for implementors and evidence for the utility of speech input for command activation in application programs.

1. Introduction

Since the advent of digital computing technology in the late 1950s, researchers have worked on developing the technology to provide voice input interfaces to computers based upon automatic speech recognition. Due to advances made in the 1960s and 1970s in the areas of digital signal processing, pattern matching and classification algorithms, and computer hardware technology, the dream of providing a speech-based input interface to the computer has become a reality, so that today, modestly priced automatic speech recognition devices are commercially available even for personal computer environments. Yet, well known problems with the current technology, such as the inadequacy of speech capture devices, speaker dependency, and restricted vocabulary size, remain significant deterrents to widespread acceptance by computer users. Despite the advances in technology, human factors research since the late 1970s has provided no conclusive evidence that automatic speech recognition is superior to conventional input devices such as keyboards and mice. However, recent studies indicate that there may be advantages to providing an additional input channel based on speech input to supplement the more common input modes.

Recently the authors conducted an experiment to demonstrate the utility of using speech input for PC-based word processing applications as a supplement to more traditional input devices such as keyboards and mice. In the experiment, speech input was used to execute word processor menu commands, while the keyboard was used for text entry and the mouse for positioning the cursor and selecting text. The authors believe that the results of the experiment provide evidence for the utility of speech input for command activation in PC and office-based applications.

1.1 Previous Research

In the early 1980s, empirical evaluations of the utility of speech input in human-computer interfaces focussed primarily on determining performance differences when speech input replaced traditional keyboard input in restricted applications. Performance measures for these early studies were usually speed and error rates, and results were typically contradictory and inconclusive.

Using a discrete word, speaker dependent system, Poock (1982) conducted a study that compared the speed and accuracy of speech input versus typed input of commands in a simulated military command and control application over a distributed network. Using a 180 word vocabulary, a recognition accuracy of 96.8% was achieved by novice users of voice input. The results showed that voice input was 17.5% faster than typed input, and that typed input of commands had 183.2% more errors than voice.

Nye (1982) reported that voice input can dramatically reduce error rates in airline baggage sorting tasks. Voice input of baggage destinations was performed with an error rate of 1% as opposed to an error rate of 10% to 40% for keyed input.

Leggett & Williams (1984) conducted an experiment to assess the performance of speech input relative to keyboard input for computer program entry and editing tasks using a language-directed program editor. Novice users of speech input completed 20-25% more of the input and editing tasks when using the keyboard as compared with speech. However, keyboard input had significantly higher error rates than speech input, and speech was just as efficient as the keyboard for issuing editing commands.

Martin (1989) cites a few other studies that received mixed results. A study by Cochran, Riley & Stuart (1980) showed speech input was slower but more accurate than typed input for entering interconnections in a circuit layout. Haller, Mutschler, & Voss (1984) performed a study that showed that voice input was slower and less accurate than keyboard input for positioning the cursor and correcting typing errors. Finally, Visick, Johnson & Long (1984) showed that for a parcel sorting task, voice input was 37% faster but was less accurate than typed input.

These earlier studies taken as a whole are not conclusive. However, the advantages of voice input over keyboard input for command activation (Poock, 1982), in applications for the handicapped (Damper, 1984), and in certain "hands busy, eyes busy" applications such as package sorting and computer aided drafting (CAD) (Martin, 1976; Nye 1982) is clear. Studies conducted by Gould (1978) and Gould *et al*. (1985) and research at IBM (Jelinek, 1985; Wylegala 1989) indicate that automatic dictation by discrete word, large vocabulary recognition systems is useful and may be practicable. However, there are indications that speech is probably more efficient than typing in tasks involving short transactions and high interaction with the computer, and less efficient for tasks that require thinking time or long transactions (Chapanis, 1975; Martin, 1989). Speech input has been shown to be beneficial for dictation of pre-recorded text (Joseph, 1989), and has great potential for providing remote information services via non-visual media (e.g., by telephone) (Aldefeld *et al*., 1980; Noyes & Frankish, 1989).

Studies performed more recently have focussed on the utility of voice input as an additional input channel in multimodal interfaces. In addition to the successful application of voice input to "hands-busy" and "eyes-busy" tasks, psychological research supports the view that people are more attentive and efficient in performing multiple tasks that are distributed across multiple input response channels of differing modes (e.g., vocal, motor etc.), since interference of tasks in the same modality decreases efficiency (Triesman & Davies, 1971; Allport *et al*., 1972; Wickens, 1980; Wickens *et al*., 1981a; Wickens *et al*., 1981b).

One of the earliest human factors studies of the utility of speech input in multimodal interfaces was performed by Martin (1989). The performance of speech input was compared with typed full word input, single key presses, and mouse clicks for entering commands in a graphic VLSI chip design package. Martin concluded that speech input was a more efficient response channel, since subjects completed 65% of the tasks when speech input was available along with the other input modes as compared to 38% when it was not available. Martin further concluded that speech improved subject efficiency because it added another input response channel.

More recently Schmandt *et al*. (1990) performed a study of the utility of using speech input to control window navigation in an X Window system while allowing keyboard and mouse input for other tasks, such as interaction with direct manipulation interfaces in application programs. There was no significant difference in speed between speech input and mouse input to navigate between exposed windows, however, speech was superior to the mouse when windows were partially or completely obscured. Further, users of speech tended to use more windows and allow more and greater degree of window overlap.

A study performed by Pausch & Leatherby (1991a) was conducted to evaluate the utility of speech input for graphical editors. An experimental group used speech input to enter commands and the mouse for pointing and selecting graphic objects, while a control group used only the mouse. Results showed a 21.23% overall time reduction when using voice input in parallel with the mouse as compared with the mouse alone. Recognition errors occurred roughly at a rate of 4%. The benefits of speech input stemmed from the reduction of cursor motion from the object being drawn to the toolbox on the far left and back. A follow-up study (Pausch & Leatherby, 1991b) measured the time taken to complete the same graphical editing tasks used in the previous study by subjects using keyboard accelerators (keystrokes which have been bound to application commands). A "novice" group of eight subjects who had not memorized the accelerator key bindings showed an average speed-up of 9.92% compared with the control group, while an "advanced" group of eight subjects who had memorized the bindings, showed an average speed-up of 14.51%.

1.2 Motivation for the Current Study

Several factors have motivated the authors to investigate the utility of speech input for word processing applications. First, word processing applications provide for a nice separation of input modes based upon the division of the primary task into subactivities, and therefore they are prime targets for multimodal interfaces. Typical word processing tasks have three basic activities that involve direct interaction with the computer. These are text entry, command execution, and direct manipulation activities such as cursor positioning and text selection. The authors believe that separate single input modes for each activity can increase efficiency. In a study conducted by Morrison *et al.* (1984), subjects switched modality from speech input to typed input while issuing text-editing commands, and found the switch of modality "disruptive." Given the lack of reliable voice activated word processing technology, text entry is best performed by keyboard. The mouse is generally accepted as being well suited to direct manipulation activities. Further, based on the human factors studies cited above, full typed input of commands (Poock, 1982; Martin 1989), single keyboard presses (Martin 1989), and accelerator keys (Pausch & Leatherby, 1991b) are not as efficient as speech-activated commands. The authors believe that the case for the utility of speech input in word processing applications relies upon its superiority over the mouse with respect to the activation of commands.

Word processing and text entry applications are also naturally "hands busy, eyes busy" applications. It is our belief that it is inconvenient and time consuming for the user to have to interrupt the typing of text or to move his/her eyes from the work in order to execute word processing commands. This is not necessary with speech activation of commands.

Further, the authors believe that user satisfaction is a key to the success of speech input in every day office applications. This motivated the authors to choose a typical office application such as word processing, to choose professionals for participation in the study, and to have them express their preferences for the input groups through a subjective questionnaire. Professionals are more likely to be serious about their preferences since the results of the study may affect their working environment.

Finally, the authors believe that they have identified a number of word processing tasks that will be more efficiently accomplished using voice activation of commands. The authors recognize that speech input may not be suitable for all word processing tasks, but they believe there are some advantages to speech input.

2. Experiment

The authors designed an experiment to evaluate the utility of using voice commands in parallel with a mouse (and keyboard for text entry) for word processing applications as compared with using the mouse alone to activate menu commands (and keyboard for text entry). In this counterbalanced design, the independent variables were the input group, either the voice group or the mouse group, and the order in which subjects used the input groups, either voice-first mouse-second, or mouse-first voice-second. The experiment consisted of four short word processing tasks, each requiring a different ratio of minimum keystrokes to minimum commands to complete the task. These ratios ranged from all commands and no keystrokes to roughly a ratio of 12 to 1. During each trial, measurements were taken of the speed required for the subject to complete each task, the number of subject errors in command activation, the number of recognition errors made by the recognition system, and, in the third task only, the number of times each subject failed to remember a short piece of text. The type of subject errors that were counted included incidences when incorrect commands were issued, when a command was issued and it should not have been, when a command should have been issued and it was not, and when an incorrect menu was selected by mouse users (e.g., "Edit" instead of "Format"). Typing and text selection errors were not counted. Each subject completed a subjective questionnaire after completing the four tasks using each input group.

The authors expected that the subjects would complete each task in less time when using voice input in parallel with the mouse than with the mouse alone. However, it was believed that the difference in time would grow less as the ratio of minimum keystrokes to minimum commands required to complete the task grew larger. No significant difference in user errors was expected across input groups or order. Finally, the authors conjectured that the subjects would prefer voice input overall for command activation; the subjects were expected to find that using voice required less effort, was more comfortable, and was faster than the mouse for the four tasks.

2.1 Subjects

For this counterbalanced experiment, a group of sixteen subjects were randomly assigned to two experimental groups distinguished by the order in which the input groups were to be used. Eight subjects, assigned to the mouse-first voice-second group, would complete each of the four tasks using the mouse input group first, and then repeat each task using the voice group. The other eight subjects, assigned to the voice-first mouse-second group, would begin with the voice and repeat the tasks using the mouse. The groups were balanced for age and gender. All subjects were computer or engineering professionals from COMSAT Laboratories and were non-novice users of the word processing application program. None of the subjects in the voice-first group had previously used the voice recognition system, and one of the subjects in the mouse-first group had experience with the voice recognition system. All the subjects had work experience using the mouse to activate commands in menu-driven applications on the Macintosh (Apple), including the word processing application used in the experiment. The majority of subjects were in their mid-twenties, fourteen out of sixteen were in the age range 21-28 years, one was in his thirties and the other in her forties. Ten of the subjects were male, and six were female.

2.2 Materials

The word processing application used for the experiment was Microsoft Word 4.0 (Microsoft Corporation). The experiment was run on a Macintosh II (Apple) using a 9.5" by 7" color monitor. We used an Articulate Systems Voice Navigator II (Articulate Systems Inc.) speaker dependent, discrete word recognition system. Each subject wore a noise cancelling head-mounted microphone and worked in a mostly quiet office environment.

The voice recognition system came with a predefined language file of command names for Word. When combined with a speaker's voice file (i.e., trained speech patterns for the commands in the language), the mapping between the speaker's utterances and the appropriate application commands is defined. The language file was edited to define or redefine mappings to many of the commands used in the experiment, using common names such as "Bullet" and "Subscript." Visual feedback was provided by the system immediately after each processed utterance was recognized. The language file command name associated with the stored speech template that most closely matched the spoken utterance was displayed in the top right corner of the screen on the main menu bar.

The Voice Navigator by default has a voice command structure for Word that is organized hierarchically like the menu system. Therefore, when using the default language file, issuing an editing command such as "cut" or "paste" requires speaking the menu title in which the command appears ("Edit" in this case) followed by the desired command. We found this arrangement to be cumbersome and ill-suited for voice input, and therefore we edited the language file so that each menu command was activated by a single voice command.

Four short word processing tasks were designed for the experiment. The first task given to the subjects could be completed using only commands activated by the voice or the mouse; no typing was required. In this task, the subject was given an unformatted document and was told to reformat the document using six predefined styles, which were "bullet", "figure", "figure-label", "text", and two section header styles, "level-zero" and "level-one." The subject used the mouse to select a portion of the text. He or she then would either use the mouse to select style commands in the "Work" menu, or would speak the appropriate command to format the selected portion of text, depending upon whether the subject was using, respectively, the mouse input group or the voice input group.

The second task required the subjects to type a short scientific formula that contained subscripted and superscripted text, bold text and Greek symbols. The ratio of minimum required keystrokes to minimum required commands was roughly 1.65 to 1. As shown in Figure 1, the text appeared at the top of the display, and the subjects typed the description below the double line. The subjects were asked to type the formula from left to right, activating the commands by voice or by mouse as needed.

The third task required the subjects to build a table of symbols using the copy, paste, up and down voice commands. A list of symbols and symbol descriptions were given to the subjects along with an empty table, as illustrated in Figure 2. The symbol list and the table could not be seen simultaneously on the display, thereby requiring the subject to page up and down (users of voice issued voice commands) while building the table. The following procedure was followed by each subject. For each symbol, the subject would select and copy the symbol, memorize the symbol description, page down to the table, paste the symbol in the table and enter the symbol description. The subject would then return to the top of the page for the next symbol, continuing in this manner until the table was full. If the subject could not remember the symbol description, the authors dictated the description to the subject as he/she typed and recorded a memory error for the task. The ratio of minimum required keystrokes to minimum required commands was roughly 5.57 to 1.

The fourth task required the subjects to type a short paragraph which contained subscripted, superscripted, italicized and bold text. The subjects were told to type the text from left to right and activate the commands when needed. This task had the highest keystroke to command ratio, roughly 12.4 to 1.

A subjective questionnaire containing fifteen questions was administered to the subjects at the end of the experiment. The first eight questions of the questionnaire are shown in Figure 3. Twelve of the questions asked the subjects to compare the voice input group with the mouse input group with regard to a number of performance and preference issues, including task completion times, error rates, response times, feedback, ease of use, comfort and preference. Two questions asked the subjects to describe what they thought were the advantages and disadvantages of voice input for word processing tasks. The subjects were also asked to give an indication of how often they would use speech input for word processing if it was available.

We explicitly avoided having subjects work from hardcopy in tasks #2 and #4 because there was a large variation in typing skill among our subjects. Few were highly skilled, and some were poor and couldn't type without looking at the keyboard, but all used word processors in their work. Further, the tasks were not designed to be wholly representative of tasks in which typists copy from existing material. We were more concerned with including typing in varying amounts, removing any source of variation in subject expertise (in typing or composition), and ensuring that there was some realistic "hands-busy eyes-busy" components to the tasks, while controlling the tasks enough to obtain roughly fixed keystroke/command ratios.

2.3 Procedure

The following procedure was followed for each subject. To begin, the subject tried on the head-mounted microphone and adjusted it so that the microphone was located approximately 1/2 inch from the corner of the subject's mouth. The subject trained the voice recognition system to recognize the following eighteen voice activated Microsoft Word commands, which were the only voice commands used in the experiment:

boldface	**paste**
bullet	**plain-style**
copy	**subscript**
down	**superscript**
figure	**symbol**
figure-label	**text**
italic	**times**
level-one	**undo**
level-zero	**up**

To train the system to recognize the subject's articulation of these commands, the subject was required to say each command at least three times. If all three instances were acceptable, the system stored the template and the user repeated the process for the next command. The enrollment process took between 3 and 6 minutes.

When training was finished, the subject was asked to perform four short tasks. Before each task, the subject was given a vocal description of what he or she needed to do to complete the task, followed by written instructions displayed on the terminal screen. In order to test that the voice commands were properly trained and to familiarize the subject with the commands and the task, the subject was asked to perform a short version of the task prior to beginning the actual task. To compensate for a potential order effect, half the subjects performed each task using the voice input group first, while the other subjects performed each task using the mouse group first. After a subject finished each task for the first time, he/she repeated the same task using the other input group. A stopwatch was used to determine the time it took the subject to complete each task. We counted the number of user errors, recognition errors, and memorization errors (for task #3 only) that occurred during the performance of each task. After completing all of the tasks using both the mouse and the voice input groups, the subject was asked to fill out a subjective questionnaire.

3. Results

Results for the experiment are in the form of task completion times, error rates, including those for user errors, recognition errors and memory errors (task #3 only), as well as subjective questionnaire scores.

3.1 Task Completion Times

The authors performed a between group analysis of the task completion times by comparing the time to complete the first trial of each task for the voice-first group versus the mouse-first group. Table 1 shows the average completion times for each task, along with the percent reduction in time to complete tasks using the voice input group as compared with the mouse group. Figure 4 shows a bar chart of the same data. On average, the subjects completed each task faster using voice input as compared with the mouse. The reduction in completion time ranged from 12.00 to 24.88 percent, and the overall reduction in task time was 18.67%. Two by two ANOVAs, for the two independent variables (order, and input group) were performed for each task, and the results are displayed in Table 2. The results showed that the reduction in task time for the voice group was significant for tasks #1 ($p<0.01$), #2 ($p<0.01$), #3 ($p<0.05$) and #4 ($p<0.01$). The order in which the tasks were performed had no significant effect on the results.

Subjects performed the tasks faster the second time when they switched input groups, and the average percent reduction was higher for all tasks. Table 3 summarizes the results for the second trials.

Throughout the experiment there were only four task trials out of 64 in which a subject completed a task faster with the mouse than with voice. The percent reductions in time for task #1 for individual subjects in the voice-first mouse-second group varied from -10.05% to 30.39%. For task #2, the percent reductions varied from -18.54% to 42.41%. For task #3, the percent reductions varied from a minimum of -12.18% to a maximum 19.32%. A minimum of 2.44% and a maximum of 26.54% was obtained for task #4. Naturally, individual reductions in speed were higher on average for subjects in the mouse-first voice-second group.

3.2 Errors

User errors related to command activation were collected for each task. A bar chart in Figure 5 compares the user errors for the first trials of each task between input groups. A two by two ANOVA was computed for the user errors, for the independent variables, order and input group, and no significant differences were found for either variable. Results of the ANOVAs are shown in Table 4. Overall, subjects made an average of 1.41 errors when using the mouse as compared with 1.08 when voice was used.

During the third task, the subjects were told to build a table of symbols by repeatedly copying the symbol and memorizing a corresponding short symbol description. The symbols were more difficult to remember when subjects used voice input as compared with the mouse. The results of the experiment show that when the voice-first group performed the task using voice input they forgot a total of 28 descriptions out of 120 compared with 13 times out of 120 for the mouse-first group using mouse input. When the voice-first group repeated the task using the mouse, they forgot only 2 descriptions. However, when the mouse-first group repeated the task using voice they forgot 27 out of 120 descriptions even though they had seen and typed the descriptions only minutes before. A two by two ANOVA was performed for independent variables, order and input. The results, shown in

Table 5, confirm that the voice input group had a significant ($p<0.01$) negative effect on memory errors as compared with the mouse-only group. Considering only the first trial of each subject group, the subjects using voice input (voice-first group) forgot an average of 3.5 descriptions while those using the mouse (mouse-first group) forgot an average of 1.63 descriptions.

The voice recognition system failed to recognize a total of 163 commands. This accounts for 6.25% of the total speech-activated commands used during the experiment. The number of recognition errors was 21 for the first task, 49 for the second task, 66 for the third task, and 27 for the fourth task. A single subject was responsible for 52 of these recognition errors. Ignoring the contribution of this subject, the recognition error rate would be 4.54%.

3.3 Subjective Questionnaire

Questions 1 through 9 of the questionnaire asked the subjects to compare the performance of the voice input group with the mouse input group using a nine point scale, as illustrated in Figure 3. A within subjects analysis of the scores for each question was performed using t-tests.

Subjective scores for the first question indicate that subjects thought that voice input was significantly faster than mouse input for tasks #1 ($t(15)=7.79$, $p<0.01$), #2 ($t(15)=6.26$, $p<0.01$), and #4 ($t(15)=5.13$, $p<0.01$). Subjects believed on average that they performed task #3 quicker when using voice input (a mean of 6.2 for voice vs. 5.1 for the mouse), however the results are not significant ($t(15)=1.56$, $p>0.05$).

The second question asked the subjects to rate the number of errors they made with each form of input. The average ratings were 3.8 for voice and 4.6 for mouse. This indicates that the subjects believed that they made fewer mistakes while using voice, however the results are not significant ($t(15)=1.62$, $p>0.05$).

In question 3, the subjects indicated that it was more difficult to remember the menu commands and their locations in the menu hierarchy when using the mouse input group than it was to remember the voice commands when using voice input. The average mouse rating was 5.2. The average voice rating was 3.7 with 1 being easier to remember and 9 being harder to remember. Again, the results were not significant ($t(15)=1.58$, $p>0.05$).

In question 4, the response time for voice was given an average of 6.1 with 9 being fastest. The response time for mouse input was given a lower average rating of 4.9. This difference was significant ($t(15)=1.78$, $p<0.05$).

The subjects thought that the feedback from the system for mouse activation of menu-commands was better than the feedback provided by the voice recognition system. In question 5, the average rating for the mouse was 6.4. It was 5.4 for voice. This difference was not significant ($t(15)=1.53$, $p>0.05$).

In question 6, the subjects indicated that the mouse required at least twice as much effort to use on average. The mouse was given an average rating of 7.1 while voice received an average rating of 3.5. This difference was significant ($t(15)=8.13$, $p<0.01$).

The subjects did not seem to believe that either input group was significantly more comfortable or relaxing to use than the other. In question 7, voice input was rated as slightly more comfortable with an average of 6.1 while the mouse had an average of 5.6 ($t(15)=0.79$, $p>0.05$).

The subjects felt that the voice input group was significantly better suited for the word processing tasks than was the mouse input group ($t(15)=4.17$, $p<0.01$). In the eighth question, the average value for the voice was 7.4, whereas it was 4.4 for the mouse.

The results for question 9 showed that the subjects significantly preferred using the voice input group over the mouse input group for the word processing tasks ($t(15)=5.35$, $p<0.01$). The rating for voice was 7.6 while the mouse rating was 4.8.

Question 10 asked the subjects to estimate how often they would use voice input while working on word processing tasks. Overall, the subjects estimated that they would use voice input 66.9% of the time. Two subjects estimated that they would use voice 100% of the time. The lowest response was a 20% estimate.

4. Discussion

There are three primary reasons for speech input being faster overall. First, when a subject wished to execute a menu command and his/her hand was not on the mouse, then he/she had to take time to locate the mouse. Often this required the subject to remove one of his/her hands from the keyboard. Further, the subject had to then locate the mouse pointer on the screen, move the mouse such that the pointer reached the desired menu (e.g., the "Edit" menu), and then traverse the menu list to reach the desired command (e.g., "cut", "copy", "paste", etc.). Even when no errors occurred, this took considerably more time than merely speaking the desired command. Six subjects volunteered on the questionnaire that it was an advantage of voice input that mouse movement for activating commands could be avoided.

Second, using the mouse to activate the menu commands required that the subject remove his/her eyes from the work, (e.g., the text on the screen, but in general it could be a piece of handwritten or typewritten text off to the side of the computer). This is necessary since the subject had to visually select the command from the menu. Time was lost due to the fact that the subject had to find his/her place again in the work after executing the command. Two subjects volunteered that speech input had the advantage of allowing them to keep their eyes concentrated on their work.

Third, using the mouse to activate menu commands required that subjects remove their hands from the keyboard. This was especially a disadvantage for typing tasks such as task #2 and task #4. Eight subjects volunteered that it was an advantage of speech input that they did not have to take their hands away from the keyboard.

The four tasks were chosen so as to give a good mix of typing and command execution. The first task required no typing and only commands. The remaining tasks had an increasing amount of typing relative to the minimum number of commands required to complete the task. For each task, the minimum number of keypresses required and the minimum number of commands required to complete each task were counted. The ratios of minimum required keypresses to minimum required commands for tasks #1, #2, #3, and #4, were found to be, respectively, 0, 1.65, 5.57, and 12.4, as shown in Table 6. A graph shown in Figure 6 shows the average percent reduction in speed between mouse users and voice users versus the ratio of minimum required keystrokes to commands. The authors expected to obtain a smooth curve showing a decrease in the reduction in speed as the ratio of keystrokes to commands increased. The unexpected low reduction value obtained for the third task suggests that the time to memorize the symbol descriptions was a significant factor that held back voice users. Considering only the other three data points, we can conclude that the percent speed up in task completion time depends upon the amount of typing involved relative to the number of commands required to complete the task. Clearly then, as expected, speech activation of commands will be superior to mouse activation of commands in command-intensive tasks as opposed to typing-intensive tasks.

For the first and second tasks, which required subjects to reformat a document, and type a scientific formula, respectively, few or no keypresses were required. The greatest decrease in task completion time for speech input users was observed for these tasks. This suggests that similar tasks involving little or no typing, (e.g., search and replace tasks, print previews, spell checks, etc.), will have similar results. Further, it is probable that tasks involving voice activation of commands deeper in the menu hierarchy, such as dialog commands for document and paragraph formatting, commands for typesetting mathematical formulae (i.e., summations, integrals, fractions, square roots, etc.), and for building and editing tables, would benefit from voice activation of commands.

It was an unexpected result that issuing commands by voice interfered with the memorization and recall of the symbol descriptions in task #3. Obviously, speaking a command interfered with the subject's short term memory. At least half of the subjects volunteered that they thought the descriptions were more difficult to memorize and recall when using voice input. One subject wrote, "in the table task it was hard to remember the description of the symbol after using the command by voice." Another wrote that the vocalization of the command "at times overrides any talking to yourself." Further, most of the subjects, when using voice input, paused momentarily after copying the symbol to memorize the description. We expected that the reduction in task time would be somewhere between that for the second and fourth tasks (e.g., about 20%). However, the reduction was much lower than expected at 12%. The time taken to concentrate on the memorization of the descriptions washed out the reduction in time that would have occurred for issuing the copy command by voice. A smaller washing out effect probably occurred for the reduction in time due to issuing the page down command by voice. The mouse users could memorize the descriptions as they moved the mouse to select the menu commands, and therefore did not lose any time. Clearly, speaking the commands interfered with short term memory. This interference is probably best explained as an interaction between two tasks of the same modality. What is unclear, however, is whether this interference will affect more complicated mental tasks such as composition and problem solving. Further study of this problem may be warranted.

As expected, the number of user errors was not dependent on the input group used. Subjects made more user errors when performing tasks #2 and #4. Inadequate feedback from the voice recognition system, combined with an underestimation of the response time were responsible for the larger number of errors made by voice input users for tasks #2 and #4 as compared with #1 or #3. Some subjects, particularly the better typists, had a tendency to type immediately after issuing the voice commands, but before feedback was given. This was further exacerbated by the fact that feedback, in the form of echoed commands on the menu bar, appeared prior to the execution of the menu commands and not afterwards. Further, the subjects seemed to expect a quicker response time for the voice activated commands. Mouse users had more difficulty with these tasks because the pull-down menus obscured the text that they were copying. They also had to move their eyes from the text in order to find the commands in the menu, and move their hands from the keyboard to locate the mouse.

Overall, the subjects reacted positively to their experience with voice input. Most subjects saw the advantage of using speech input in what they thought were "hands busy, eyes busy" tasks. One subject wrote that voice input "reduces hand/mouse movement...[and] enables user to keep hands on keyboard, which speeds up processing time." Others wrote that speech input "doesn't require [you] take [your] hands off of the keyboard to use the mouse as often," you do "not have to interrupt typing to select from the menu bar," speech causes "less distraction from typing [and] allows more concentration on reading and typing," and "you don't have to take your eyes off the text you are typing and in some cases you don't have to take your fingers off the keyboard." By way of generally positive comments, subjects wrote that the advantages of speech are that it "is faster and more accurate," it doesn't require "memorization of menus," "complicated actions are easier to perform," it "prevents memorization of accelerators" and "in frequent tasks such as copy and paste, or changing styles ... voice is better."

Subjects wrote that the disadvantages of speech input were "commands are not always reliably recognized," "side conversations can accidentally [cause] changes in your document," "erroneous commands can cause a little disorientation, but 'undo' helps -- also deactivating voice would help," "I don't talk exactly the same at all times," "what if I have a cold?," and "having everyone 'vocalizing' their word processing up and down the hall would be nerve-racking." This supports the view that recognition error rates, background noise and user attitude are important considerations in the success of speech input devices. Some thought the feedback provided by the speech recognition system was inadequate. One subject writes "since it's not always reliable it's hard to determine if it actually worked. Only sometimes can I get a glimpse of the menu bar item flashing." Response time was a concern of one subject. There was a "slower response time for the voice command input." One subject seemed to dislike using speech input. This subject had the largest number of recognition errors, that is, 33 for task #3, 10 for task #4, and 52 overall. We believe that inconsistent speaking style and attitude may have been the cause, and that with more training the subject could have achieved acceptable performance levels, but we can not be certain.

5. Conclusions

The authors conducted an experiment to measure the utility of using a voice recognition system to allow speech-activated commands in word processing applications as compared with using the mouse to activate menu commands. Overall, averaging the results of the first trials for all four tasks, subjects using the voice first were 18.67% faster in completing the tasks than subjects using the mouse first. This result was obtained despite the fact that, one, the subjects were novice users of speech input, two, they received at most one or two minutes training before each task, not counting the enrollment time, and three, the subjects had all previously used the mouse to execute menu commands in the word processing application as part of their work. Intermittent and experienced users of speech input may achieve greater reductions in task completion time. This conjecture is supported by the results for the second trials (Table 3), which show a greater decrease in task times for speech input users over the first trials. The error rates due to user mistakes were roughly the same for both input groups, and recognition errors, averaged over all the tasks, occurred for 6.25% of the speech-activated commands. As expected, the subjects reacted positively to using voice input, however they had concerns about recognition accuracy, response time, and feedback. One unexpected result was that issuing speech input commands significantly interferes with short term memory.

The results suggest that speech input for command activation provides improved performance over mouse activation of commands in word processing applications, particularly for tasks that are command intensive or that require formatting of text as it is entered, as in the scientific formula task (#2) and the long typing task (#4). Although our subject pool was relatively small (16 in all), based on our results it appears that professionals using computers in an office environment may find speech input helpful in their everyday word processing tasks. These results may be generalized to other applications in which there is a clear separation of command execution and other activities that engage the hands or eyes, such as text entry, direct manipulation of text or graphics, and other activities outside the computer domain such as occur in package sorting and inspection tasks. Speech input is not likely to replace other modes of input, but seems to have a useful place along with other types of input in a multimodal interface.

5.1 Impact for Practitioners

The study indicates that at least some people (including the majority of those who participated in the study) would prefer to use speech input for executing commands in word processing applications. However, designers of speech recognition systems should make attempts to provide adequate feedback so that users can recognize and prevent substitution, false acceptance and recognition failure errors. Clearly, better feedback would have improved the performance of the subjects in this study.

Improvements in response time would also decrease the number of user errors, especially in typing tasks similar to tasks #2 and #4. It is apparent that some users are anxious to begin typing immediately after issuing the voice command so that they can avoid typing interruptions. However, with the system used in this experiment, some subjects had to learn to wait for feedback from the command before typing.

Improvements in speech capture devices are needed as well. The head-mounted microphone works well, but it is often too much trouble to use (especially for short transactions) and uncomfortable to wear. A desk-top microphone and a lapel microphone were used in pilot studies. The desk top microphone, our first choice, proved to be very sensitive to background noise and dependent upon the position and orientation of the speaker, which was difficult to keep constant since there was no positional dependency between the subject and the microphone. The lapel microphone eliminated some of the problems due to the lack of position dependency, but since the microphone could not be positioned close enough to the user's mouth, recognition performance was poor.

The authors expect that the utility of speech input for command activation in application software will be improved when developers begin to design their products so that they can maximize the benefits of speech input. This may result in improvements in feedback, response time, user control of command definition and editing, and so on, that would help offset the serious drawbacks of current technological restrictions in speech capture devices, speaker dependency and vocabulary size.

5.2 Suggestions for Future Research

An in depth study of how issuing voice commands affect short term memory and task performance is needed. Such a study should focus on the performance of tasks that range from "simple" memorization tasks to more complicated problem-solving tasks so as to determine the extent of the interference and the range of situations under which such interference occurs or is significant. It would be interesting to see how this interference with memory relates to previous studies of the effects of interference between tasks of the same modality.

A study of the effect of display size on the performance of speech-activated commands versus mouse activation of menu commands might be useful. It might also be useful to determine the practical limits on vocabulary size by looking at how vocabulary size affects overall user performance, user memory and command recall, and system performance. A long term study of expert users of speech input in an integrated applications environment (e.g., PC-based automated office environment) might give more solid user satisfaction and performance data.

Acknowledgements

The authors would like to thank the men and women from the Systems Development Division at COMSAT Laboratories who participated in this experiment. Their contributions to this experiment were invaluable and are greatly appreciated.

References

Aldefeld, B., Rabiner, L. R., Rosenberg, A. E., & Wilpon, J. G. (1980). Automated directory listing retrieval system based on isolated word recognition. *Proceedings of the IEEE, 68,* 11, 1364-1379.

Allport, D. A., Antonis, B., & Reynolds, P. (1972). On the division of attention: a disproof of the single-channel hypothesis. *Quarterly Journal of Experimental Psychology, 24* , 225-235.

Chapanis, A. (1975). Interactive human communication. *Scientific American, 232*, 36-42.

Cochran, D. J., Riley, M. W., & Stewart, L. A. (1980). An evaluation of the strengths, weaknesses and uses of voice input devices. *Proceedings of the Human Factors Society--24th Annual Meeting*. Los Angeles.

Damper, R. I. (1984). Voice-input aids for the physically handicapped. *International Journal of Man-Machine Studies 21*, 541-553.

Gould, J. D. (1978). How experts dictate. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 4, 648-661.

Gould, J. D., Conti, J., & Hovanyecz, T. (1985). Composing letters with a simulated listening typewriter. *Communications of the ACM 26,* 4, 295-308.

Haller, R., Mutschler, H. & Voss M. (1984). Comparison of input devices for correction of typing errors in office systems. *Proceedings of INTERACT '84, First IFIP Conference on Human-Computer Interaction,* London.

Jelinek, F. (November 1985). The development of an experimental discrete dictation recognizer. *Proceedings of the IEEE, 73,* 11, 1616-1624.

Joseph, R. (1989). Large vocabulary voice-to-text systems for medical reporting. *Speech Technology*, April, 49-51.

Leggett, J., & Williams, G. (1984). An empirical investigation of voice as an input modality for computer programming. *International Journal of Man-Machine Studies 21*, 493-520.

Martin, T. B. (1976). Practical applications of voice input to machines. *Proceedings of the IEEE, 64,* 4, 487-501.

Martin, G. L. (1989). The utility of speech input in user-computer interfaces. *International Journal of Man-Machine Studies 30* , 355-375.

Morrison, D. L., Green, T. R. G., Shaw, A. C., & Payne, S. J. (1984). Speech-controlled text-editing: effects of input modality and of command structure. *International Journal of Man-Machine Studies 21*, 49-63.

Noyes, M. N., & Frankish, C. R. (1989). A review of speech recognition applications in the office. *Behavior and Technology, 8,* 6, 475-486.

Nye, J. M. (1982). Human factors analysis of speech recognition systems. *Speech Technology, 1,* 2, 50-57.

Pausch, R., & Leatherby, J. H. (1991a). A study comparing mouse-only input vs. mouse-plus-voice input for a graphical editor. *Journal of American Voice Input/Output Society, 9*, 2.

Pausch, R., & Leatherby, J. H. (1991b). Voice input vs. keyboard accelerators: a user study. *Proceedings of the AVIOS '91 Voice I/O Systems Applications Conference*, 9-14.

Poock, G. K. (1982). Voice recognition boosts command terminal throughput. *Speech Technology, 1,* 2, 36-39.

Schmandt, C., Ackerman, M. S., & Hindus, D. (1990). Augmenting a window system with speech input. *IEEE Computer* , 50-56.

Triesman, A., & Davies, A. (1971). Divided attention to ear and eye, in *Attention and Performance, Vol IV* , 101-117.

Visick, D., Johnson, P. & Long, J. (1984). The use of simple speech recognisers in industrial applications. *Proceedings of INTERACT '84, First IFIP Conference on Human-Computer Interaction,* London.

Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson & R. Pew, Eds. *Attention and Performance VIII.* New York: Erlbaum.

Wickens, C. D., Mountford, S. J., & Schreiner, W. (1981a). Multiple resources, task-hemispheric integrity, and individual differences in time-sharing. *Human Factors, 23*, 211-230.

Wickens, C. D., Vidulich, M., Sandry, D. & Schiflett, S. (1981b). Factors influencing the performance advantage of speech technology. *Proceedings of the Human Factors Society--25th Annual Meeting,* Rochester, NY.

Wylegala W. (1989). A 20,000-word recognizer based on statistical evaluation methods. *Speech Technology*, April, 16-18.

$$d^2(\mathbf{r}_v) / dt^2 = - \mu \, \mathbf{r}_v / r_v^3 + \mu_m (- \mathbf{d}_{mv} / d_{mv}^3 - \mathbf{r}_m / r_m^3)$$

==

(Subjects entered the text here)

Figure 1. Task #2, Typing a Scientific Formula

μ = Earth gravitational constant

R_e = Earth equatorial radius

L_s = luminosity of the Sun

A_s = satellite surface area

$\mathbf{r}_s$ = satellite position vector

$\mathbf{v}_s$ = satellite velocity vector

η = satellite surface reflectance

α_{GM} = Greenwich hour angle

ω_s = satellite angular velocity

m_s = satellite mass

ρ = slant range

R_p = Earth polar radius

ϕ_g = geographic latitude

θ_g = geographic longitude

C_{ds} = drag coefficient of the satellite

Symbol	**Description**

Figure 2. Task #3, Building a Table of Symbols

1. You were able to complete the following tasks faster for which input group?

a) The Formatting Task (#1):

Mouse-only: Slowest <- 1 2 3 4 5 6 7 8 9 -> Fastest
Voice-plus-Mouse: Slowest <- 1 2 3 4 5 6 7 8 9 -> Fastest

b) The Scientific Formula Task (#2):

Mouse-only: Slowest <- 1 2 3 4 5 6 7 8 9 -> Fastest
Voice-plus-Mouse: Slowest <- 1 2 3 4 5 6 7 8 9 -> Fastest

c) The Table of Symbols Task (#3):

Mouse-only: Slowest <- 1 2 3 4 5 6 7 8 9 -> Fastest
Voice-plus-Mouse: Slowest <- 1 2 3 4 5 6 7 8 9 -> Fastest

d) The Long Typing Task (#4):

Mouse-only: Slowest <- 1 2 3 4 5 6 7 8 9 -> Fastest
Voice-plus-Mouse: Slowest <- 1 2 3 4 5 6 7 8 9 -> Fastest

2. With which input group did you feel you made the least mistakes?

Mouse-only: Least <- 1 2 3 4 5 6 7 8 9 -> Most
Voice-plus-Mouse: Least <- 1 2 3 4 5 6 7 8 9 -> Most

3. For which input group were the commands and their locations within the menu (if mouse-only) easiest to recall?

Mouse-only: Easiest <- 1 2 3 4 5 6 7 8 9 -> Hardest
Voice-plus-Mouse: Easiest <- 1 2 3 4 5 6 7 8 9 -> Hardest

4. How would you rate the response time for voice command input as compared with the mouse-menu command input?

Mouse-only: Slowest <- 1 2 3 4 5 6 7 8 9 -> Fastest
Voice-plus-Mouse: Slowest <- 1 2 3 4 5 6 7 8 9 -> Fastest

5. How would you rate the feedback provided for voice command input as compared with the mouse-menu command input?

Mouse-only: Worst <- 1 2 3 4 5 6 7 8 9 -> Best
Voice-plus-Mouse: Worst <- 1 2 3 4 5 6 7 8 9 -> Best

Figure 3. Subjective Questionnaire, Questions 1 - 5

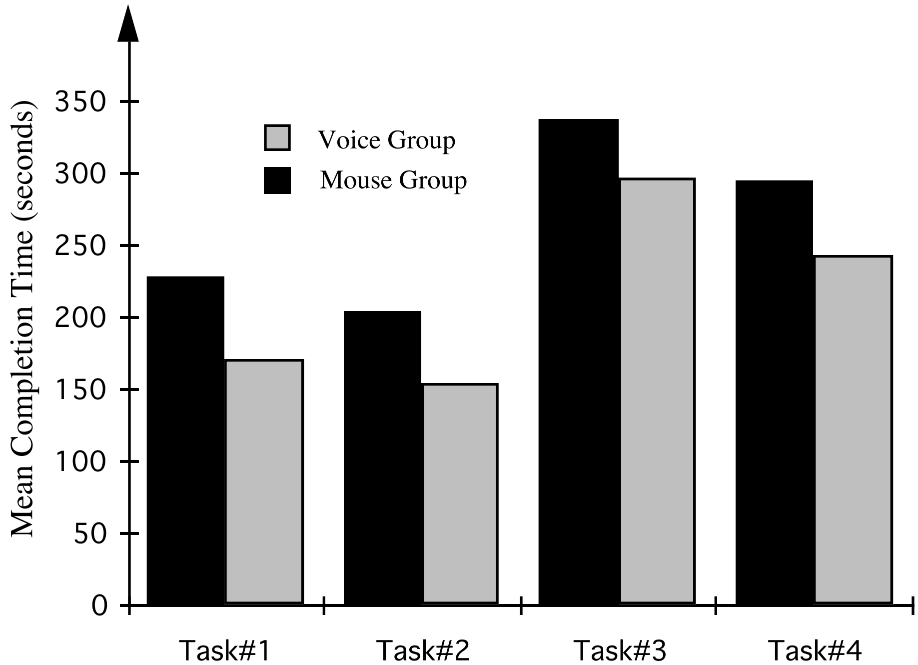

Figure 4. Comparison of Mean Task Completion Times

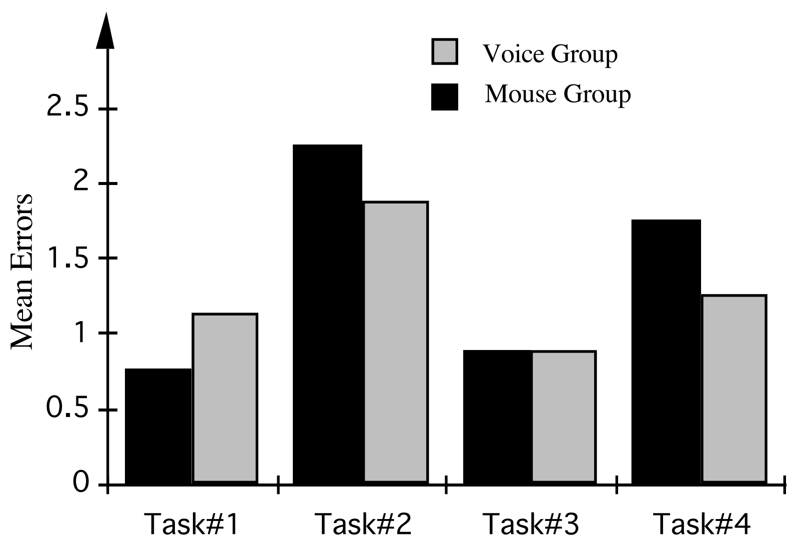

Figure 5. Comparison of Mean User Errors, First Trials, n=8

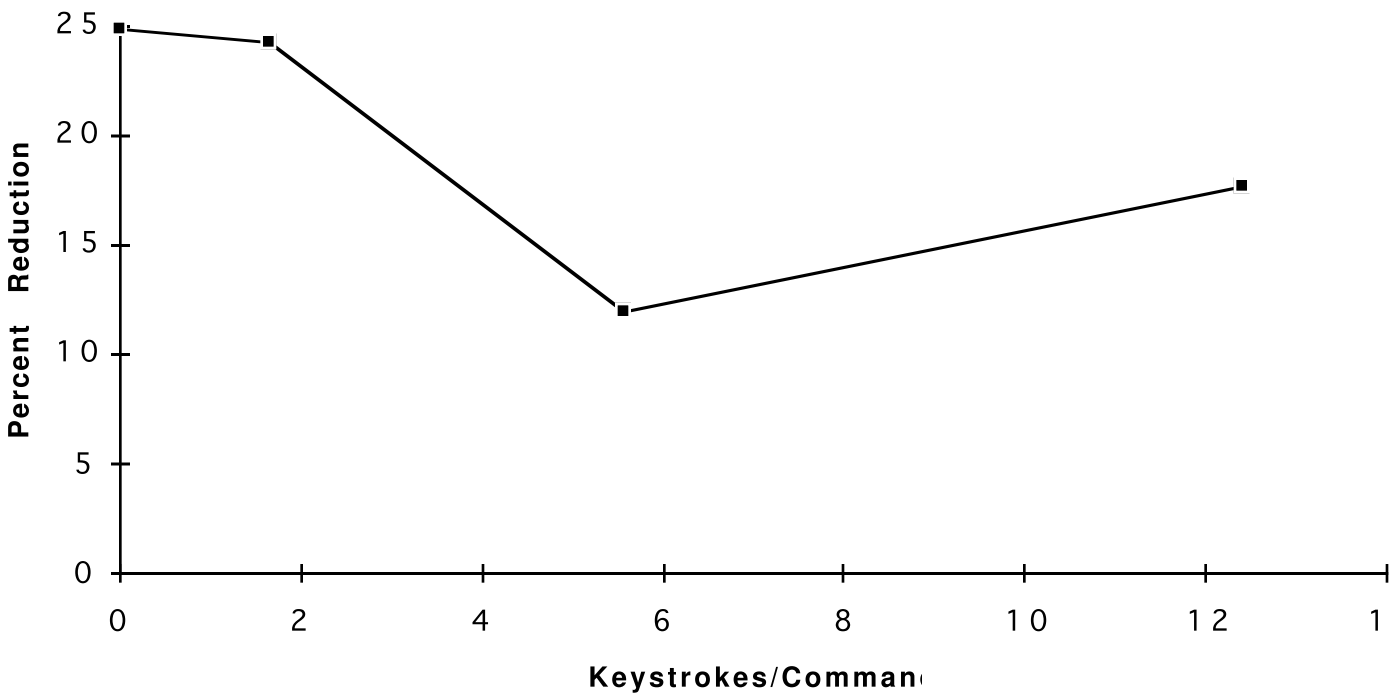

Figure 6. Graph of Percent Reduction in Time vs. Ratio of Minimum Required Keystrokes to Minimum Required Commands

Table 1. Average Task Times and Percent Reduction Between Groups for First Trials (n=8, standard deviation in parentheses)

Task Number	Avg. Time Mouse-First (sec.)	Avg. Time Voice-First (sec.)	Percent Reduction
1	228.6 (41.1)	171.8 (30.9)	24.88
2	203.6 (38.7)	154.4 (23.8)	24.19
3	337.5 (52.5)	297.0 (29.0)	12.00
4	296.0 (72.6)	243.6 (41.6)	17.69
Total	266.4	216.7	18.67

Table 2. F Values for 2x2 ANOVA for Times (Order x Input); df = (1,28), $F_{0.01}$ = 7.64, $F_{0.05}$ = 4.20

2x2 ANOVA F Measures	Task #1	Task #2	Task #3	Task #4
F_{input}	24.00	18.5	5.30	7.98
F_{order}	0.0534	0.282	0.00148	0.256
$F_{order\ x\ input}$	5.80	0.0549	0.831	0.880

Table 3. Average Task Times and Percent Reduction Between Groups for Second Trials (n=8, standard deviation in parentheses)

Task Number	Avg. Time Mouse-Second (sec.)	Avg. Time Voice-Second (sec.)	Percent Reduction
1	199.4 (19.6)	147.6 (20.3)	25.96
2	207.5 (44.9)	144.4 (26.5)	30.42
3	321.9 (45.2)	280.0 (57.6)	13.01
4	286.2 (47.6)	211.0 (70.8)	26.29
Total	253.8	195.8	22.86

Table 4. F Values for 2x2 ANOVA for User Errors (Order x Input); df = (1,28), $F_{0.01}$ = 7.64, $F_{0.05}$ = 4.20

2x2 ANOVA F Measures	Task #1	Task #2	Task #3	Task #4
F_{input}	0.0323	1.52	0.612	0.223
F_{order}	1.58	0.24	0.612	0.223
$F_{order\ x\ input}$	2.61	0.24	0.0680	0.0558

Table 5. F values for 2x2 ANOVA of Total Memory Errors (order x input); df = (1,28), $F_{0.01}$ = 7.64, $F_{0.05}$ = 4.20

2x2 ANOVA F Measures	F Values for User Errors
F_{input}	16.4
F_{order}	1.03
$F_{order\ x\ input}$	1.48

Table 6. Ratios of Minimum Keystrokes to Minimum Commands

Task	Minimum Required Commands	Minimum Required Keystrokes	Ratio of Keystrokes to Commands
#1	34	0	0
#2	35	58	1.65
#3	60	334	5.57
#4	34	422	12.4