SPEECH VERSUS MANUAL CONTROL OF CAMERA FUNCTIONS DURING A TELEROBOTIC TASK

John M. Bierschwale, Carlos E. Sampaio, Mark A. Stuart, and Randy L. Smith
Lockheed Engineering and Sciences Company

Research directed by Jay Legendre, Manager, Remote Operator Interaction Lab, NASA JSC.

INTRODUCTION

Telerobotic workstations will play a major role in the assembly of Space Station *Freedom* and later in construction and maintenance in space. Successful completion of these activities will require consideration of many different activities integral to effective operation: operating the manipulator, controlling remote lighting, camera selection and operation, image processing, as well as monitoring system information on all of these activities.

Of these activities, the vision (camera viewing) system is particularly important. During many tasks where a direct view is not possible, cameras will be the user's only form of visual feedback. If the vision system is manually controlled and both hands are busy during the performance of a dynamic task, it will require reorientation of the hands and eyes between the manipulator controls, vision system controls, and view of the remote worksite. Allocating some or all of the control of vision system components to voice input may lessen the workload of the operator, reduce movement time, and ultimately improve performance. Voice input is currently being considered for this as well as other applications by NASA.

Very few studies are found in the literature that investigate the use of voice input for camera control. The only study that was found (Bejczy, Dotson, Brown, and Lewis, 1982) was relevant in that it investigated voice input for camera control of the Remote Manipulator System (RMS) and payload bay cameras used on the Space Shuttle. Although statistical analyses were not presented, voice input was found to be 10% slower across four subjects.

The philosophy of the present investigation differs in that subjects were not constrained to current RMS control panel terminology and organization. Subjects used words from a vocabulary sheet developed in a previous study (Bierschwale, Sampaio, Stuart, and Smith, 1989) to construct camera commands to accomplish a telerobotic task. The subjects' vocabulary preferences are presented elsewhere (Bierschwale, et al., 1989).

It is important to consider current terminology so that personnel are not forced to learn new jargon. However, the use of voice input was not considered in the development and selection of the current terminology and switch labels. Choice of vocabulary is very important in terms of recognizer performance and user acceptance. Successful vocabulary design (ultimately the human machine interface design) will most readily be achieved by considering the recognition qualities of the commands and cognitive relationship between the commands and their respective actions.

A potential problem with voice control of cameras may be verbalizing the directions to move the cameras. Many people have difficulty when providing verbal directions. An example would be saying "left" when "right" is meant. Indeed, this cognitive difficulty when verbalizing directions has been noted with voice control of cursor movement while editing text (Murray, Praag, and Gilfoil, 1983; and Bierschwale, 1987).

Identification of critical issues such as this early in the design phase will allow for more effective implementation of a voice

commanded camera control system. In more general terms, one report (Simpson, McCauley, Roland, Ruth, and Williges, 1985, p. 120) found that, historically, "projects designed from inception to incorporate a voice interactive system had a greater probability of success than when the capability was added to an existing system." By understanding the differences between the two modes of input, a more effective utilization can be made of both voice and manual input.

The objectives of this study are as follows: (1) optimize the vocabulary used in a voice input system from a Human Factors perspective, (2) perform a comparison between voice and manual input in terms of various performance parameters, and (3) identify factors that differ between voice and manual control of camera functions.

METHOD

SUBJECTS

Eight volunteer subjects were selected to participate in this evaluation. These subjects were partitioned into the following two groups: an experienced group of four subjects who were familiar with telerobotic tasks and workstations and an inexperienced group of four who were not familiar with these concepts.

APPARATUS

Testing took place in the Man-Systems Telerobotic Laboratory (MSTL) located at the NASA Johnson Space Center. A Kraft manipulator slaved to a replica master controller was used to perform a remote telerobotic task. The task selected for this study was a generic pick and placement task. This task required a high degree of visual inspection and dextrous manipulation. The tasksite is depicted in Figure 1.

Two 4-inch tall and two 10-inch tall tiers were placed on a semicircular taskboard in front of the Kraft manipulator. Three task pieces were placed on the lower left-hand tier and three were placed on the upper left-hand tier. On the right-hand side, four receptacles were placed on the upper and lower tiers (two receptacles per tier). The task consisted of locating, grasping, transporting, and depositing each of four task pieces into the correct receptacle. In addition to the required manipulation, subjects had to move cameras, adjust lens parameters, and select views to successfully complete the task. During the task, subjects were instructed which task piece and receptacle were involved.

Two cameras equipped with remote pan, tilt, zoom, focus, and iris controls provided the operator with two oblique views of the worksite (i.e., approximately 45 degrees above the horizontal plane with one displaced 45 degrees to the left and the other camera 45 degrees to the right). A fixed-focus camera provided a "bird's-eye" view of the entire work area looking down at a 45-degree angle on the worksite from above the task. The two oblique views were input to a 21-inch monitor where only one view could be shown at a time. The "bird's-eye" view was continuously displayed on a 9-inch monitor positioned atop the larger monitor.

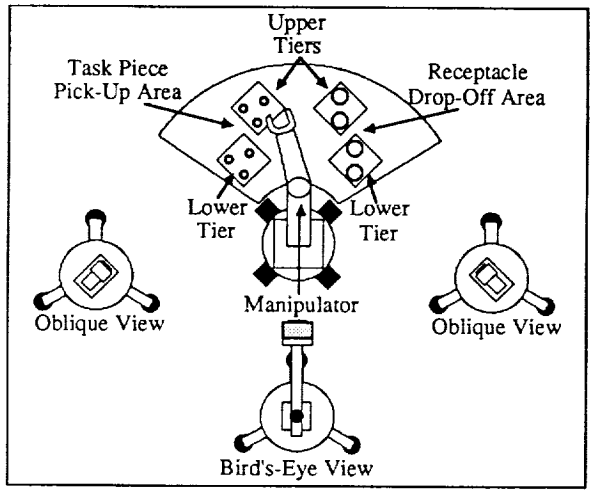

Figure 1. Overhead view of remote work site.

The left oblique view showed the task pieces and the surrounding area. The right oblique view showed the box and the surrounding area. The "bird's-eye" view showed the entire work area. Taskpieces were aligned such that the left oblique view was required to read their

markings while the right oblique view was required to read the receptacles' markings.

A practice task (using direct view) was devised so subjects could become familiar with the manipulator controls, camera views, and kinesthetics of the arm movements and positions that they would be using later during data collection. The practice task used the identical views, taskboard, tier placement, and similar task objectives.

During use of manual input, the camera controls were placed directly in front of the subjects. Subjects were required to use their right hand to operate both the manipulator and camera controls. This required halting the manipulator to operate the cameras. This was a simulation of a hands-busy scenario where voice input might aid performance.

A vocabulary list containing stereotypical words determined in a previous evaluation (Bierschwale et al., 1989) was used for voice input. A separate vocabulary sheet was used for each subject and the words were randomly listed under each icon (descriptive of the camera function) to avoid any possible list order effect.

In order for the control system to be flexible enough to accommodate the various word combinations, an experimental approach was used that has been referred to as the "Wizard of Oz" method. This is often used in user-computer interaction research and is summarized in Green and Wei-Haas (1985). For this evaluation, a "wizard" carried out the actions of a speech recognizer. This method has been used before with voice input research. One study (Casali, Dryden, and Williges, 1988) used a wizard recognizer to evaluate the effects of recognition accuracy and vocabulary size on performance.

The "wizard" was situated at the camera controls out of the field-of-view behind and to the right of the subject. When voice input was used, the "wizard" wore a headset which allowed him to screen out external noise and concentrate on the commands issued by the subject through a microphone.

VARIABLES

Three independent variables with two levels each were studied in this evaluation: input modality (voice or manual), level of experience (experienced or inexperienced), and administration order (voice followed by manual input or manual followed by voice input). Experience level and administration order were between-subjects variables while input modality was a within-subjects variable.

Dependent variables consisted of task completion time, number of camera commands, and errors. Scaled question and questionnaire responses were also collected.

PROCEDURES

At the beginning of the evaluation, subjects were provided with a brief explanation of the purpose of the study. Each subject received instruction on the use of the manipulator controls that would be needed to perform the manipulation tasks.

Following performance of the practice task (using direct view), a videotape was used to illustrate the different camera and lens movements that would be available on the two adjustable cameras. The investigator used deliberate wording when pointing to the corresponding icons on either the camera control panel (manual) or the vocabulary sheet (voice), so as not to bias any subject's selection of vocal commands. Prior to each of the two conditions, subjects were instructed on the use of the respective camera controls. When using manual input, a template with descriptive icons illustrating the functions was placed over the controls so that the subjects would not be biased in their vocabulary selection (for voice input) by using the listed labels. These same icons were used on the vocabulary sheet for voice input.

The subjects' view of the task was then obstructed so that they had to rely totally on the camera views. Each subject performed two sessions under both conditions (voice and manual input). The first was a practice session using an abbreviated version of the task with

the second being the complete task. This practice also allowed subjects, while using voice input, to become familiar with the designated words and select the few they might prefer to use. Administration order was counterbalanced with half of the subjects using voice input first and half using manual input first. Additionally, to avoid any memorization of task requirements, different locations and task piece selections were used for each of the four sessions (one practice and data collection session per condition).

Following the practice session, the data collection session was conducted. Subjects were instructed to work quickly while making as few errors as possible. If an error occurred, the taskpieces and receptacles were placed in the configuration present prior to the error and the subject repeated the trial. While setup time was not recorded, repetition of the trial was included in the completion time. The video images used to perform the task (excluding the "bird's-eye" view) were recorded along with audio input from the subjects' headset.

Following completion of the data collection session for each condition, subjects completed a questionnaire. The procedure for the second condition progressed in the same manner. After testing was finished, another questionnaire, involving comparison of the two modalities, was completed by the subjects.

RESULTS AND DISCUSSION

PERFORMANCE DATA

Analysis of Variance (ANOVA) results are presented for the task completion times, number of commands, and errors. Table 1 presents the group means for each of these measures.

An ANOVA run on the task completion times found that voice input was significantly slower than manual input for controlling cameras in this task ($F(1,4) = 19.80$, $p < .05$).

In order to allow for a direct comparison of the number of camera manipulations, the voice commands were tallied such that a single command consisted of both activating and stopping the movement (actually two voice commands issued). An ANOVA run on the number of commands that were used found that significantly more commands were used with manual than voice input ($F(1,4) = 10.34$, $p < .05$).

It was expected that more manual commands would be used since people tend to "bump" manual controls and set things up perfectly. With voice input, subjects tended to accept coarse adjustments because of the difficulties imposed by the system lag time and lack of variable rate control. If examined in conjunction with the task completion times, it is seen that subjects used more time to execute fewer commands with voice input. It may very well be that using voice input to control the cameras resulted in more cognitive difficulty associated with each command which could result in more errors. On the other hand, assuming a constant error rate, the greater number of commands given with manual input would increase the probability that an error will occur. If this effect exists, this is a system trade-off that will need to be evaluated.

TABLE 1.

Group means for performance measures.

	Voice input		Manual input	
	Exp.	Inexp.	Exp.	Inexp.
Completion Time (Minutes)	12.58	15.46	10.94	12.41
Commands *	90.30	104.50	114.00	130.30
Manipulation Errors	.75	.75	1.00	1.00
Focusing Error Rates (percent)	30.50	32.80	50.00	37.00
* Does not include extra commands resulting from directional errors.				

It was hypothesized that fewer manipulator errors would occur with voice control since this would allow the subjects to keep their eyes on the screen and avoid interruption of the task. However, the results of an ANOVA

show that there was not a statistically significant difference in the number of manipulation errors between modalities. The makeup of the task was such that few errors were committed with either modality.

A directional error consisted of moving one direction when one wanted to move in the opposite direction. Very few directional errors were observed across the functions except when focusing the cameras. More errors were made when using manual control to focus the cameras than when using voice control. However, results of an ANOVA revealed no significant difference in focusing error rates between the two input modalities.

The probable reason for the high focusing error rates was that the task required zooming the focal length back and forth and the subject would usually guess which directional command would bring the picture into focus. Possible reasons why somewhat higher error rates occurred during manual input were that subjects tended to perform more commands, as was previously mentioned, and were more likely to attempt to bring the picture into exact focus. With voice input, focusing was difficult due to the sensitivity of the focusing operation and the system lag time. Subjects would often accept a less than perfect image.

SUBJECTIVE RESPONSES

The following types of questions were asked concerning the two input modalities: scaled questions, open-ended questions, and yes/no questions that allowed the subject to elaborate. Analysis revealed no real differences in preference between the two modalities of input and two experience groups across all of the questions for this task. However, similar subjective comments, concerning advantages and disadvantages of the two modalities for performance of this task, were frequently made across many of the questions and are summarized in Table 2.

When subjects were asked what telerobotic workstation functions they would recommend allocating to voice input, the following applications were given: selecting or moving cameras, controlling lights, halting the manipulator arm, setting the manipulator grip lock, changing modes, and panning and tilting only. For the most part, these applications are of a discrete nature that minimize the disadvantages of voice input listed in Table 2.

TABLE 2.

Advantages and disadvantages of voice-operated camera control.

VOICE INPUT	
ADVANTAGES	DISADVANTAGES
Hands and eyes free	Cognitive difficulty verbalizing commands/directions
Good for single, gross movements while hands are occupied	System lag time
Possibility for simultaneous camera/ manipulator control	Two step start-stop process
	Can't perform two camera movements at once
MANUAL INPUT	
ADVANTAGES	DISADVANTAGES
Finer positioning than voice input	Diverting eyes and hands from telerobotic task to adjust camera controls
Less mental load than voice input	
Quicker system response time	

CONCLUSION

This investigation has evaluated the voice-commanded camera control concept. For this particular task, total voice control of continuous and discrete camera functions was significantly slower than manual control. There was no significant difference between voice and manual input for several types of errors. There was not a clear trend in subjective preference (across several questions) of camera command input modality. Task performance, in terms of both accuracy and speed, was very similar across both levels of experience.

One problem that emerged was that numerous focusing errors (30-50%) were observed across both groups and modalities. For tasks as dynamic as this, development of an

autofocusing system is highly recommended to avoid operator frustration and inefficiency.

The fundamental advantage that voice input had over manual input, as mentioned by both groups of subjects, was that it allowed the hands and eyes to be free to do other tasks.

Unfortunately, voice input of camera controls also resulted in cognitive difficulty when verbally transcribing movements, specifying the correct directions, and stopping movements. The advantage of manual input was that it allowed precise positioning. The applications that subjects suggested for voice input at a telerobotic workstation were of a discrete control nature.

Most of the problems seem to be associated with the movement processes. While each distinct movement (zoom, pan, tilt, etc.) was not directly compared across both modalities, subjective comments indicate that the problem is a fundamental one of verbal control of a spatial motor task. The study by Bejczy, et al. (1982) also stated that controlling camera movement was troublesome for the subjects.

The results of this investigation indicate that using voice input for control of discrete types of camera operations (selecting cameras, multiplexing, and selecting rates) could aid performance in a telerobotic task. Control of continuous camera functions by voice input is not recommended.

A combination of voice input and manual input for control of camera movements would take advantage of the best aspects of each of the control modalities. Future studies should evaluate alternate methods of controlling camera movements. Some examples are: (1) a hand-controller mounted joystick whose function is selected (camera pitch, camera roll, zoom, focus, and iris control) by voice and controlled manually, which would save panel space by only requiring one control for each or all of the cameras, (2) activating movements by voice and stopping them manually using a switch on the hand controller, and (3) use discrete levels of zoom, focus, and iris (Level 1, 2, 3, 4, etc.) and discrete movements of cameras (perhaps angular, as in pan right 30°, 60°, etc.). Other modalities of input such as an eye tracking device or head-slaved camera control device should also be investigated.

These results were achieved with a particular task, manipulator, and camera control system. A voice recognizer simulation was used that had the advantage of 100% recognition and the possible disadvantage of slower response time. An actual voice recognizer will not perform this well. With decreasing recognition rates, several things will probably occur (although it is difficult to precisely quantify the magnitude of the effects). For example, one study (Casali et al., 1988) found a 17% increase in completion time for a data entry task when recognition rate dropped from 99 to 95% and a 50% increase in completion time when the rate dropped from 99 to 91%. It was also found that each lowered level of recognition produced a significant decline in subjective acceptance of the system.

Different tasks and control systems might produce different results, although it is believed that the trends discussed in this report are applicable across a wide variety of telerobotic tasks. Thus, it is contended that the results will have immediate application to the design of the telerobotic workstations.

ACKNOWLEDGEMENTS

Support for this investigation was provided by the National Aeronautics and Space Administration through Contract NAS 9-17900 to Lockheed Engineering and Sciences Company.

REFERENCES

1. Bejczy, A. K., Dotson, R. S., Brown, J. W., and Lewis, J. L. (1982). Voice control of the space shuttle video system. In *Proceedings of the 17th Annual Conference on Manual Control* (pp. 627-640). Pasadena, CA: Jet Propulsion Laboratory and California Institute of Technology.

2. Bierschwale, J. M. (1987). *Speech versus keying commands in a text-editing task.* Unpublished master's thesis, Texas A&M University, College Station, TX.

3. Bierschwale, J. M., Sampaio, C. E., Stuart, M. A., and Smith, R. L. *Development of a vocabulary for voice input of camera controls during a telerobotic task.* (NASA Tech. Report JSC - 23706). Houston, TX: NASA Lyndon B. Johnson Space Center.

4. Casali, S. P., Dryden, R. D., and Williges, B. H. (1988). The effects of recognition accuracy and vocabulary size of a speech recognition system on task performance and user acceptance. In *Proceedings of the Human Factors Society 29th Annual Meeting* (pp. 232-236). Santa Monica, CA: Human Factors Society.

5. Green, P., and Wei-Haas, L. (1985). The rapid development of user interfaces: experience with the Wizard of Oz method. In *Proceedings of the Human Factors Society 29th Annual Meeting* (pp.470-474). Santa Monica, CA: Human Factors Society.

6. Murray, J. T., Van Praag, J., and Gilfoil, D. (1983). Voice versus keyboard control of cursor motion. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 103). Santa Monica, CA: Human Factors Society.

7. Simpson, C. A., McCauley, M. E., Roland, E. F., Ruth, J. C., and Williges, B. H. (1985). System design for speech recognition and generation. *Human Factors, 27,* 115-141.