Technical Memorandum 10-91

THE EFFECTS OF USER'S TRAINING ON THE PERFORMANCE OF AN AUTOMATIC SPEECH RECOGNIZER FOR A SELF-PACED TASK

Christopher C. Smyth

April 1991
AMCMS Code 612716H700011

U.S. ARMY HUMAN ENGINEERING LABORATORY
Aberdeen Proving Ground, Maryland

®Aydin is a registered trademark of Aydin Controls.

®DEC, UNIBUS, VAX, VMS are registered trademarks of Digital Equipment Corporation.

®SAS statistical analysis computer package is a registered trademark of SAS Institute Incorporated.

®Verbex series 4000 is a registered trademark of Voice Industries.

Destroy this report when no longer needed.
Do not return it to the originator.

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Use of trade names in this report does not constitute an official endorsement or approval of the use of such commercial products.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Technical Memorandum 10-91	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Human Engineering Laboratory 6b. OFFICE SYMBOL (If applicable) SLCHE	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Aberdeen Proving Ground, MD 21005-5001	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 6.27.16 / PROJECT NO. 1L162716AH70 / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
The Effects of User's Training on The Performance of an Automatic Speech Recognizer for a Self-Paced Task

12. PERSONAL AUTHOR(S)
Smyth, Christopher C.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM _____ TO _____	1991, April	81

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	automatic speech recognition, air defense, training, aviation, tactical displays
01	02		
01	0301		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The results of a recent experiment concerning the effects of training on the performance of subjects using the automatic speech recognizer are reported. Over a 5-day period, 20 military enlisted grade male subjects were trained and tested in using a connected speech (speaker-dependent) machine automatic speech recognizer in a self-paced task controlling a generic tactical display by voice command.

Experimental results show that a majority of the subjects had little difficulty with the automatic speech recognizer and that for these subjects training produced only a slight improvement in recognizer performance. These subjects performed at a high machine recognition rate. However, during the first session, a large minority (35%) of the subjects had difficulty training their speech to be machine recognizable. These subjects required at least two training sessions to perform the task at their best ability, and even after they were trained, their performance never reached the performance level of the other subjects.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	Unclassified
22. NAME OF RESPONSIBLE INDIVIDUAL Technical Reports Office	22b. TELEPHONE (Include Area Code) (301) 278-4478 22c. OFFICE SYMBOL SLCHE-SS-TSB

DD Form 1473, JUN 86 *Previous editions are obsolete.*

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

AMCMS Code 612716H700011 Technical Memorandum 10-91

THE EFFECTS OF USER'S TRAINING ON THE PERFORMANCE OF AN AUTOMATIC SPEECH RECOGNIZER FOR A SELF-PACED TASK

Christopher C. Smyth

April 1991

APPROVED: _____
JOHN D. WEISZ
Director
Human Engineering Laboratory

U.S. ARMY HUMAN ENGINEERING LABORATORY
Aberdeen Proving Ground, Maryland 21005-5001

ACKNOWLEDGMENTS

The author wishes to thank Mr. Alan M. Poston, Research, Development, and Engineering Coordinator, Field Support Division, U.S. Army Human Engineering Laboratory (HEL), for his discussions during his assignment as aviation team leader about the purpose and methodology of this study. Because of Mr. Poston's suggestion, the scope of the investigation was expanded to include speech retention and task variability.

The need for a study about training effects was realized by the author during discussions with Mr. Jerome H. Howie, Ford Motor Company, and Ms. Kathleen A. Christ, Corporate Technology Office, U.S. Army Laboratory Command, about applications to automotive quality control at the 1987 International Speech Technology Conference in London, England. At that time, Ms. Christ was an HEL industrial engineer working on applications of speech recognizers to Army aviation.

The author appreciates the support, advice, and guidance of Mr. Frank J. Malkin, formally aviation team leader, Aviation and Air Defense Division, HEL. Mr. Malkin's pioneering investigations in HEL about automatic speech recognizer performance in support of the light helicopter experimental (LHX) and advanced rotary technology integration (ARTI) program set the standards for further research in this area.

CONTENTS

FIGURES

TABLES

EXECUTIVE SUMMARY

The results of a recent experiment concerning the effects of training on the performance of subjects using the automatic speech recognizer are reported. In the future, speech recognition technology is expected to have a dramatic impact on the design of Army vehicle cockpits and command and control display consoles by allowing displays to be controlled by voice commands alone. For this reason, it is important to establish limitations of present technology.

Over a 5-day period, 20 military enlisted grade male subjects were trained and tested in using a connected speech (speaker-dependent) machine automatic speech recognizer during a self-paced task, controlling a generic tactical display by voice command. Using short phrases as commands, the subjects extracted and modified data about track symbols displayed on a screen. The subjects were trained in using the automatic speech recognizer, data entry tasks, and speech commands, and they practiced with the automatic speech recognizer before being tested. The number of machine recognitions was recorded during the test. This experiment determined the effects on the automatic speech recognizer performance of (a) subject training, (b) template enrollment procedure, (c) speech retention, and (d) task variability.

The experiment was conducted as follows: During the first 3 days, the subjects were enrolled daily on the automatic speech recognizer and trained in the data entry task before testing. Video tapes of the task were used during training. The enrollment procedure was necessary to establish the speech templates that the speaker intended to use with the automatic speech recognizer. The subjects made an enrollment update on the automatic speech recognizer at the start of the fourth day and were allowed to practice before testing. On the fifth day, the subjects operated the automatic speech recognizer with the templates from the previous day. The practice and test sessions of the fifth day were followed by a test of the effects of changes in the display update rate.

Experimental results show that a majority of the subjects had little difficulty with the automatic speech recognizer and that for these subjects, training produced only a slight improvement in recognizer performance. These subjects performed at a high machine recognition rate. However, during the first session, a large minority (35%) of the subjects had difficulty training their speech to be machine recognizable. These subjects required at least two training sessions to perform the task at their best ability, and even after they were trained, their performance never reached the performance level of the other subjects.

The average performance after full training was 89% recognition accuracy. This is consistent with the results of previous experiments at the U.S. Army Human Engineering Laboratory (HEL) and other laboratories with operational environments.

The subjects' performances degraded unless their speech templates were enrolled on the machine daily before use. Subjects could not retain the speech patterns necessary to maintain performance without daily enrollment training. This was also true after the 3-day training period was completed. The use of enrollment training in place of full enrollment, following the initial three-session subject training period, did not affect performance.

One disturbing result was the occasional occurrence of consecutive runs of multiple nonrecognitions during testing. This resulted from subtle voice changes of which the subject was apparently unaware. These runs of multiple nonrecognitions were a major source of errors, which occurred primarily among subjects who needed training during the training phase and among other subjects during the speech-retention tests. Such nonrecognitions rarely occurred after full training with daily enrollment.

Finally, the changes in the display update rate significantly increased the variation in recognizer performance (at least during the first few test runs of that session). This was apparently because of the initial need for practice during the changed test conditions.

The results suggest that the performance of a speaker-dependent machine system can be influenced by subtle variations in speech from those of the machine enrollment training set. Furthermore, the persistence of a human speaker in maintaining consistent speech may be determined by his experience in giving verbal commands and by the task to be performed. There is evidence that the older subjects with higher military rank and those in combat arms performed at a higher machine recognizer rate. Certainly, the results suggest that a large minority of the military population will need training before they can attain functional performance with machine recognizers.

Apparently, some of the Army population will need to repeat daily enrollment training before use to achieve reasonable performance. It should be noted that the enrollment procedure for a speaker-dependent (connected speech) automatic speech recognizer can be lengthy, and in some cases, can require an hour of training for a large command vocabulary.

For these reasons, the author recommends careful consideration of the speech patterns of the intended military user population until more robust, speaker-independent, machine automatic speech recognizers are developed. Since technology is continually improving, HEL supports research in technology applications for Army combat vehicles and command and control centers.

THE EFFECTS OF USER'S TRAINING ON THE PERFORMANCE OF AN AUTOMATIC SPEECH RECOGNIZER FOR A SELF-PACED TASK

INTRODUCTION AND BACKGROUND

Speech recognition technology will have a dramatic impact on the design of future Army vehicl. crew station designs, aircraft cockpits, and command and control display consoles by allowing voice command entries to free the soldier's hands and eyes for other tasks. Army helicopter pilots have found that the use of automatic speech recognizers reduces the need for looking at cockpit displays and using manual actions during the performance of certain tasks. Using automatic speech recognizers interferes minimally with pilotage and therefore contributes to the flight stability needed during nap-of-the-earth (NOE) flight.

However, the verbal utterances used with the automatic speech recognizer increase the time to perform a task. Furthermore, most present-day automatic speech recognizers are speaker dependent and their performance is affected by the pilot's condition. The pilots' acceptance of automatic speech recognizers in the cockpit is mixed, varying from enthusiastic to tolerant. For these reasons, the application of present-day automatic speech recognizers in a helicopter cockpit has been limited to non-time-critical tasks in which the manual backup has high visual and motor demands. An example of a successful application is the control of radio communications.

Considering the potential usefulness of automatic speech recognizers, it is important that one understands some limitations of present technology. A major limitation may be the long training and adaptation period required by some users of automatic speech recognizers to reach functional performance. The need for an adaptation period would explain the differences in performance reported for some human factors studies as opposed to industrial applications. For example, a 1987 U.S. Army Human Engineering Laboratory (HEL) study (Smyth, Denny, & Dotson, 1987) resulted in a 76.22% recognition accuracy for a speaker-dependent, connected word, automatic speech recognizer employed for data entry on tactical displays. This result is less than the 86% performance rate reported in another human factors study (Malkin & Christ, 1985) of a similar nature in a helicopter simulator.

These results sharply contrast those reported by the automotive and aircraft industries for applications of automatic speech recognizers. In particular, the automotive industry reported that recognition accuracies from 97% through 99% are commonly attained at assembly line inspection sites (Howie, 1987).

The most likely explanation for the discrepancy in these results is the long training and adaptation period of several weeks used by industry to prepare its workers for using automatic speech recognizers. Therefore, the primary purpose of this study is to investigate the effects of an adaptation period on training. However, the impact of other factors on machine recognition performance must be understood for proper control of test error variance.

The major factors affecting machine recognition performance (Simpson, McCawley, Roland, Ruth, & Williges, 1985) are (a) the user's motivation, (b) the user's natural speech patterns and experience in voice control, (c) the user's physical state, (d) automatic speech recognizer hardware, (e)

microphone stability, (f) enrollment procedure, (g) task vocabulary, (h) task stress, (i) task prompts and recognizer feedback, and as mentioned previously, (j) the user's training. Some factors, their possible impact on performance, and examples taken from human factors and industry studies are reviewed.

Motivation

The users of an automatic speech recognizer for assembly line inspection in the automotive industry are highly motivated, experienced quality control inspectors who have been selected because of their past performance to help develop an automatic speech recognition system for automotive inspection lines. Furthermore, they can see how the system benefits their task since they no longer need to manually record the results on a checklist. The computer, based on their voice entries, does the recording for them.

Vocabulary

The vocabulary and the frequency of occurrence of certain words influence the recognition accuracy. In the automotive inspection process, the most commonly used phrase should be "no fault" or a suitable replacement because of the low occurrence of manufacturing errors. In contrast, the occurrence of task words or phrases in a military system would be more evenly distributed throughout the recognizer vocabulary because of the diversity of tasks performed.

Prompts and Feedbacks

The modality of the task prompt and the recognizer feedback to the user may influence recognizer accuracy. In the automotive inspection process, the task prompt is provided by the user to the system; the automatic speech recognizer is always receptive for data entry. The feedback is auditory either as a reinforcing "beep" for successful recognition or a series of computer-generated synthetic speech questions asking the user about entry performed. These questions are a type of user on-the-spot training since they reinforce monotonic voice which can be used for the successful data entry into an automatic speech recognizer.

In the 1987 HEL study (Smyth et al.), the task prompt was a visual cue on a computer-driven tactical display, while the recognizer feedback to data entry was a reinforcing beep for successful recognition or no response for nonrecognition. The task feedback to the data entry was a visual change of the tactical display.

An example of the effect of feedback on task performance was provided by this study. The visual change in the tactical display was a response that the subject should have been checking. However, the recognizer feedback isolated the subject from the tactical display response. The subjects tended to channel their attention to the auditory senses and waited for the reinforcing beep, and during this process, failed to check the resulting changes in the tactical display. Consequently, nonrecognitions by the automatic speech recognizer and the accompanying erroneous changes in the display were not realized until late in the task, thereby increasing the subject's anxiety.

Enrollment Experience

There are several causes of variation between the enrollment patterns and the task patterns for the automatic speech recognizer. First, inexperienced users may have their voice utterances out of sequence with the automatic speech recognizer prompts during a poorly supervised enrollment session. This is easily done since the enrollment procedure generated by most automatic speech recognizers is based on the monotonous repetition of the voice commands in a noise-free room. Furthermore, the differences between the enrollment and the task environments can lead to different speech patterns. An inexperienced user will tend to be sedated during enrollment but excited during task performance.

Physical

Other causes affecting recognizer performance are physical. A shift in the microphone position on the subject's head can cause a deterioration in recognizer performance. The shape of a subject's nasal cavity can change because of congestion during the onset of a cold or upper respiratory infection, thereby changing speech patterns. Stimulants, such as coffee or physical exercise, can change the speech delivery rate.

Task-Related Stress

Other factors affecting recognizer performance are task related. Changes in the task can lead to emotional and physical changes in subjects causing consequential changes in their speech patterns. For example, time constraints during task performance can influence recognizer accuracy. As time to perform a task becomes shorter, the subjects must accelerate their verbal response which changes the pitch components of the utterance. For this reason, the speech is changed from the enrollment patterns and reduces recognizer accuracy.

Certainly, there is evidence that the task stress generated by time constraints can cause a decrement in voice recognizer performance. Several studies (French, 1983; Martin & Poock, 1984; Poock & Martin, 1984) have investigated the effects of emotional and perceptual motor stress on voice recognizer accuracy. Conclusions of these studies show that task stress induced by time constraints can cause recognition performance to decrease.

Interestingly, Simonov and Frolov (1973), in their study of using formant structure for a measurement of emotional stress and attention state in pilots and cosmonauts, reported a strong relation between heart rate and voice changes during changing work load.

Speech Patterns

A dominant factor appears to be the difference in speech patterns, particularly the natural consistency in voice patterns among subjects. Subjects with relaxed, deeply robust voice patterns do better than those with higher pitched voices. This is shown by the gender differences as males with low pitched voices tend to do better than females with higher pitched voices. The same is true for subjects with the more pronounced northern accents as opposed to those with southern accents of North America. Experience has shown that some subjects, when giving voice commands, naturally maintain a sufficiently consistent voice pattern and have a high rate of recognizer

performance. In contrast, other subjects tend to vary their voice patterns to the extent that the automatic speech recognizer is unable to determine their utterances from moment to moment during the task. These subjects then try different variations until the original enrollment pattern is repeated. This difference is influenced by training and prior experience. Commissioned and noncommissioned officers are trained and experienced in giving orders, flight officers are trained in radio communications, and inexperienced enlisted military soldiers are trained to follow orders.

Training

Finally, the influence of recognizer performance and user's training with the automatic speech recognizer during the task is discussed. The training on the automobile inspection line, using a automatic speech recognizer consists of three sequential stages. First, the user's voice pattern is enrolled on the automatic speech recognizer using job-related phrases. Second, a brief training period is conducted during which the job is taught using the automatic speech recognizer. Third, the user is allowed an adaptation period of one to several weeks during which he daily enrolls on the automatic speech recognizer and then uses the device on the job. During this last period, he learns to adapt his voice to the automatic speech recognizer in the work environment.

In contrast, the training in the 1987 HEL study (Smyth et al.) was abridged to just the first two stages conducted during one period. The subjects were enrolled on the machine and then given brief training about the task with the machine. The test followed after the subjects took a short rest period, and no adaptation period was provided. Most subjects thought that the device was difficult to control by voice input. An adaptation period between task training and the test may have improved their performance.

The adaptation period gives the user time to integrate the automatic speech recognizer with the task. The user will learn the voice commands and the enrollment procedure for the automatic speech recognizer. Apparently, the users realize the speech pattern intensity and pitch that they will use with each voice command during participation in the task and apply these speech patterns during the enrollment. The users set up their speech patterns in the automatic speech recognizer during the enrollment period before task application. For a robustly voiced speaker, the automatic speech recognizer works best if the task voice commands have the same speech patterns as those used during enrollment. If the task voice commands do not sound enough like those used during enrollment, the voice commands will be rejected by the automatic speech recognizer and the user cannot perform the tasks.

One facet of adaptation training is the number of training periods needed to reach an asymptote in recognizer performance. Essentially, this is the number of training periods needed to learn how to maintain speech patterns consistently enough to operate the automatic speech recognizer when users perform the task. The users are assumed to learn a sequence of pitch and intensity patterns to perform the task with voice commands.

In addition to the effects of training length and adaptation period, other training questions are

Will using enrollment training instead of full enrollment (following establishment of speaker templates) influence performance?

How long can the subject retain an asymptotic performance level following training without further enrollment? In other words, how long does it take for the learned speech patterns necessary to operate the automatic speech recognizer to reach extinction?

Can the subject maintain an asymptotic performance level during changing task conditions? That is, can the subject maintain the needed consistency in the speech patterns as the task conditions are changed? In other words, how robust are the learned speech patterns to changes in task conditions?

OBJECTIVES

The purpose of this study is to determine the effects of user training on the recognition accuracy of an automatic speech recognizer for a self-paced task, controlling a generic tactical display by voice commands. The subjects extract and modify data about track symbols displayed on the screen using short phrases as commands. In particular, the purpose of this study is to determine whether (a) a subject training period of several days will increase machine recognizer performance to an asymptotic level, (b) substitution of machine enrollment training for full machine template enrollment will influence performance following completion of subject training, (c) subjects can retain the machine performance without further enrollment following their training completion, and (d) subjects can maintain the performance during time-varying test conditions.

METHODOLOGY

The subjects were tested with the voice command portion of the tactical display test conducted by HEL in 1987 (Smyth et al.). The effects of other factors on recognizer performance were controlled as much as experimentally possible to reduce error variance. Only male subjects were studied because of differences in performance between male and female voices. The subjects were given as much assistance during the training sessions as possible. A video film of the tactical display task was shown to the subjects when the first training session began. When the next training session began, subjects saw video recordings of themselves performing the test, and recordings were discussed with them. The subjects were enrolled on the automatic speech recognizer at the start of each training session. The enrollment sessions were rigidly supervised to ensure proper matching between utterances and prompts. The voice commands used during the test were short phrases; these phrases were the only speech encouraged during the test, except for the track query task response which was not processed by the automatic speech recognizer. The experimenter maintained control of environmental noise. Subjects were advised regarding proper conduct considered to be supportive of the test. A head support system was used to hold the microphone in place.

This section covers (a) apparatus, (b) test subjects, (c) experimental design, and (d) training and test procedures.

Apparatus

The following apparatus and equipment were used during this experiment:

1. DEC VAX® 11/780 host computer consisting of a central processing unit, a floating point accelerator, and 2.75 megabytes of memory. The computer has UNIBUS® adapter interfaces to an Aydin® display processing unit, real time clock, and communications ports to VT100/220 terminals. The VAX virtual memory system (VMS®) operating system supports the FORTRAN language in real time simulation of military systems. The VMS language provides the priority, scheduling, process creation and control, real time event-driven response, and high speed, interprocess communications essential for real time simulation of complicated systems.

2. An Aydin graphics system (model 5216 display computer) providing a 1024- by 1024-pixel resolution. The system has five memory planes which can generate 16 simultaneous colors with an overlay for alphanumerics. The memory bus controller-processor controls vector and character generation and permits pixel loading from the host computer at 800 pixels per second. The refresh memory modules provide a 1024- by 1024- by 5-pixel storage resolution for the video output. Interface to the host computer is by a parallel DR11-W direct memory access UNIBUS.

3. An Aydin model 8026 color graphics, video monitor driven by the Aydin raster scan graphics system (model 5216 display computer) providing a 1024- by 1024-pixel resolution. The monitor is a 19-inch diagonal (15.5 by 11 inches), high resolution, red-green-blue (RGB) color monitor for use with the Aydin graphics system.

4. A VERBEX® Series 4000 connected word speech recognition system that is a speaker-dependent, connected speech, voice data entry peripheral that enables the entry of predefined strings of words without pauses.

5. A noise-cancelling microphone with a custom-made head support system for stability.

6. Panasonic color video camera, WV-3400, with portable VHS video cassette recorder (VCR), NV-8420, and AC adapter, NV-858, allowing both video and audio recordings.

7. A large screen (25-inch diagonal) television monitor with VCR tape player.

8. A VT100 computer terminal used to control the test program.

The interfacing of the devices to the host computer, the computer processes including the display driver, and the display concept are described as follows:

Interfacing

Figure 1 shows the experimental console and test apparatus. The experimental console holds the Aydin raster scan display screen and operator console shelf. The tactical display was shown on the raster scan display positioned between the desk and eye level. The display console shelf was at desk height. The console was designed in accordance with Sections 5.7.5 and

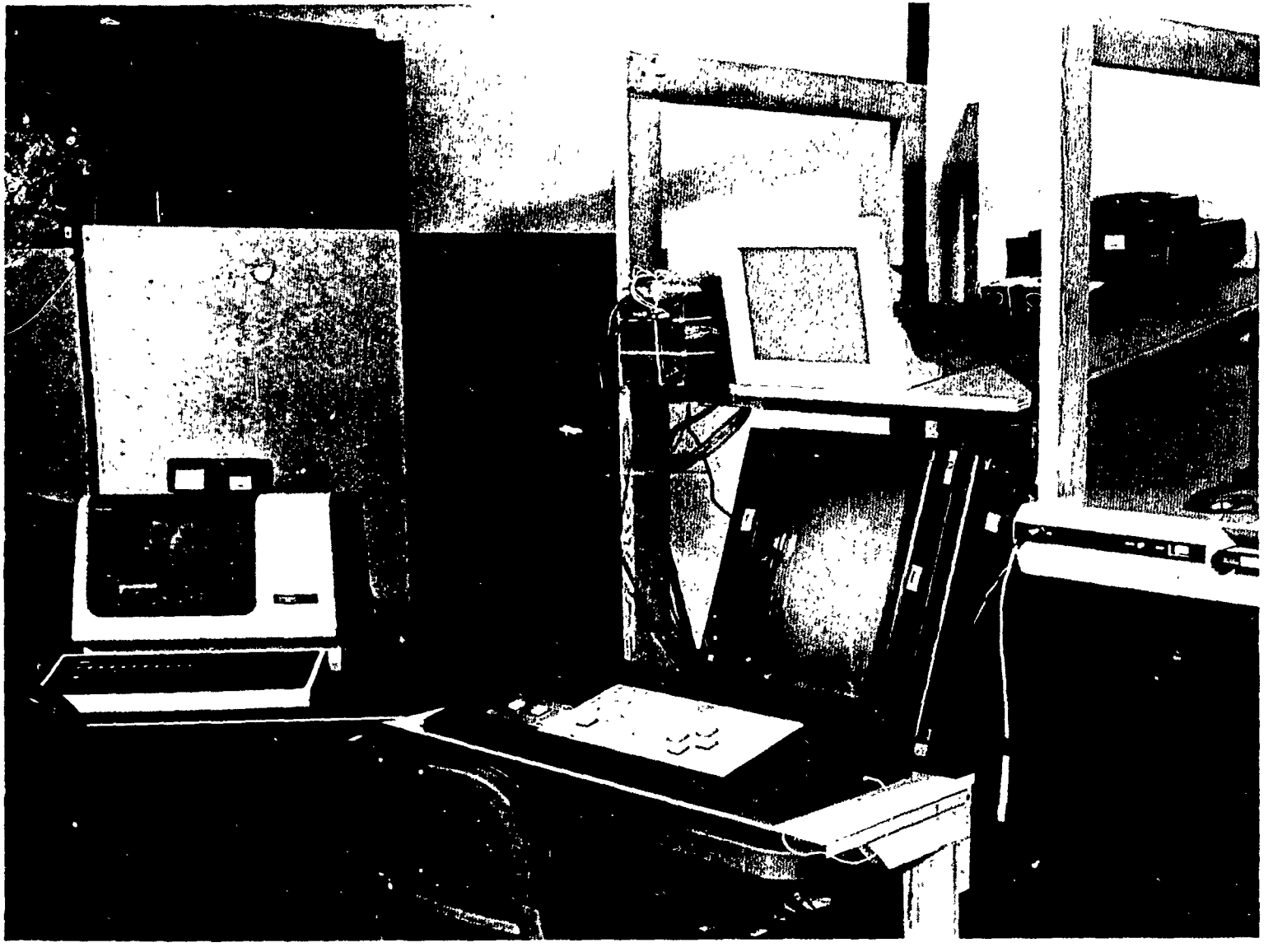

Figure 1. Test station.

5.7.6 of MIL-STD-1472C (DoD, 1981). The Verbex speech recognition system is to the right of the console; the VT100 used by the experimenter is to the left. The VCR (not shown) is positioned to the right rear of the subject to record his actions, utterances, and the tactical display during the test.

Process Control

The DEC VAX 11/780 computer is the process controller for the test. All equipment was interfaced to the computer. The Verbex was interfaced by an RS232 port. Separate computer program processors service the Verbex and display driver; the routines communicate through a global common area.

A separate process drives the Aydin monitor showing the tactical display of Figure 2. The test scenario track symbols are updated once every second on the tactical display. The voice entries, causing changes in the display in a selected track symbol or a menu, are serviced immediately for user feedback. Entries not corresponding to a selected item cause an error message to be displayed. Voice entries that are not recognized as a reference template cause the VT100 to be momentarily beeped as feedback to the subject that a mis-recognition has occurred.

The use of the VMS operating system is necessary for real time programming of complicated configurations if the response times of the different devices are to be reduced to match the sensitivity of a human subject. The VMS operating system allows the execution of different subprocesses servicing the different devices. The processes run independently during system level control but exchange data through event flags and global common areas. The subprocesses can be controlled by various system level services to schedule the processing of events.

Display Concept

The characteristics of the tactical display are presented in Table 1. The display is divided into the instruction area, the graphics display area, and the menu display area. Figures 2 through 5 show the display presentations for the data request and identification report tasks using the voice input modality.

The instruction area is at the top of the display. Figure 2 shows an example of the data request instruction message, while Figure 3 shows an example of the identification report instruction message used during this study. The graphics display area is at the center of the screen, below the instruction area. The tactical graphics area shows a real time scenario with positions of the air tracks updated every second. This area is dedicated to the status of the air picture about the host aircraft and contains 15 dynamic track symbols. Only one of the track symbols is task related and has predetermined flight characteristics. A different target track was selected for each task for every trial in a counterbalanced manner so that all flight characteristics were represented in all tasks and input modalities.

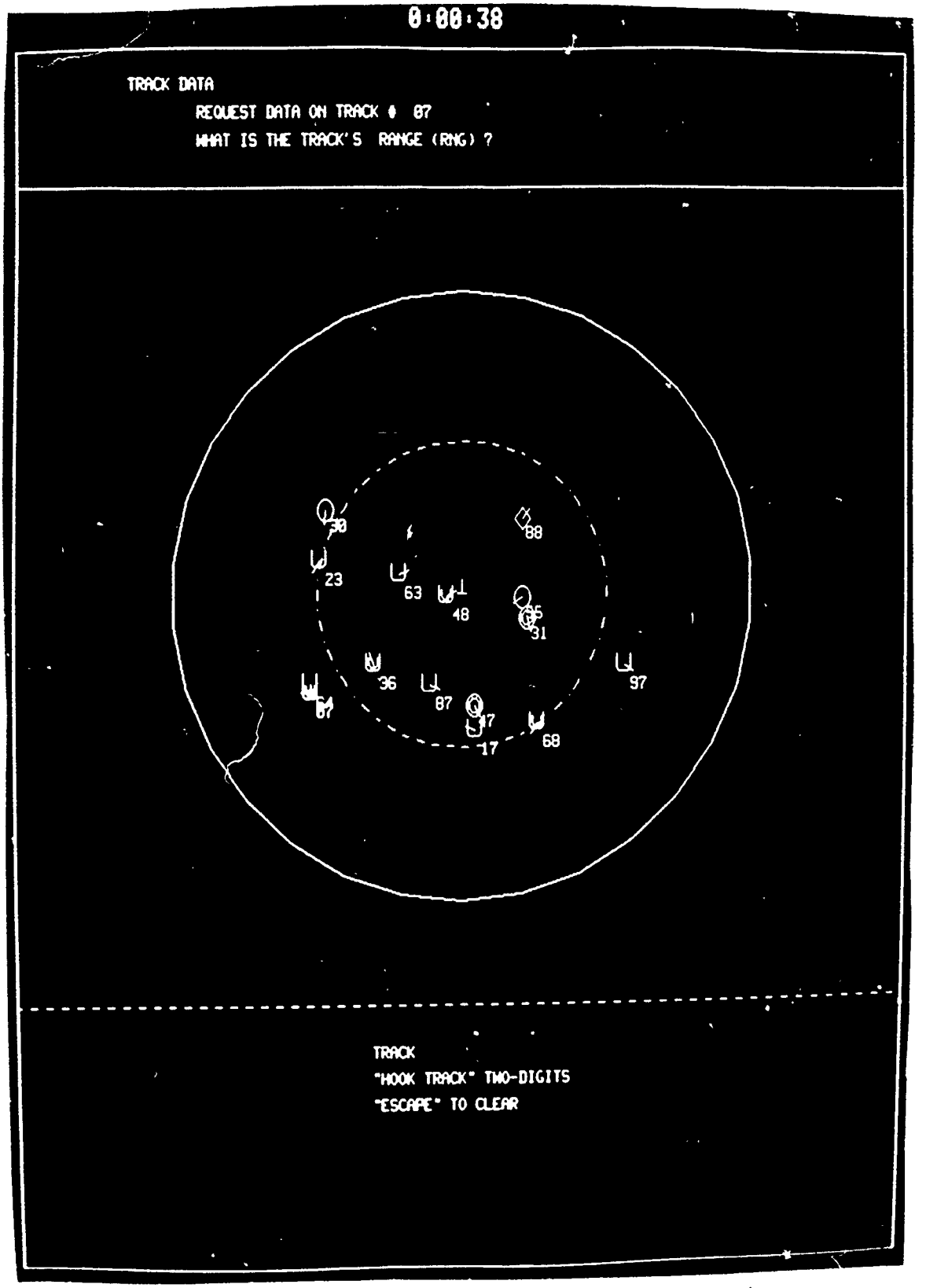

Figure 2. Track selection menu.

Table 1

Display Screen Size and Characteristics

Display element	Characteristics	Value
Aircraft symbol	Size	16 rasters square (0.228 in^2)
Fire unit symbol	Size	10 rasters square (0.142 in^2)
Range rings	Radius	
Inner ring		1.5 inches
Outer ring		3.0 inches
Alphanumeric		
Characters	Size	
Aircraft numbers		5 x 7 rasters (0.10 inch)
Menu characters		7 x 9 rasters (0.13 inch)
Screen size		
	Width	
Overall		900 rasters (10.0 inches)
	Length	
Overall		925 rasters (13.0 inches)
Instruction area		150 rasters (1.5 inches)
Graphics area		600 rasters (8.5 inches)
Menu/sub-menu area		174 rasters (3.0 inches)

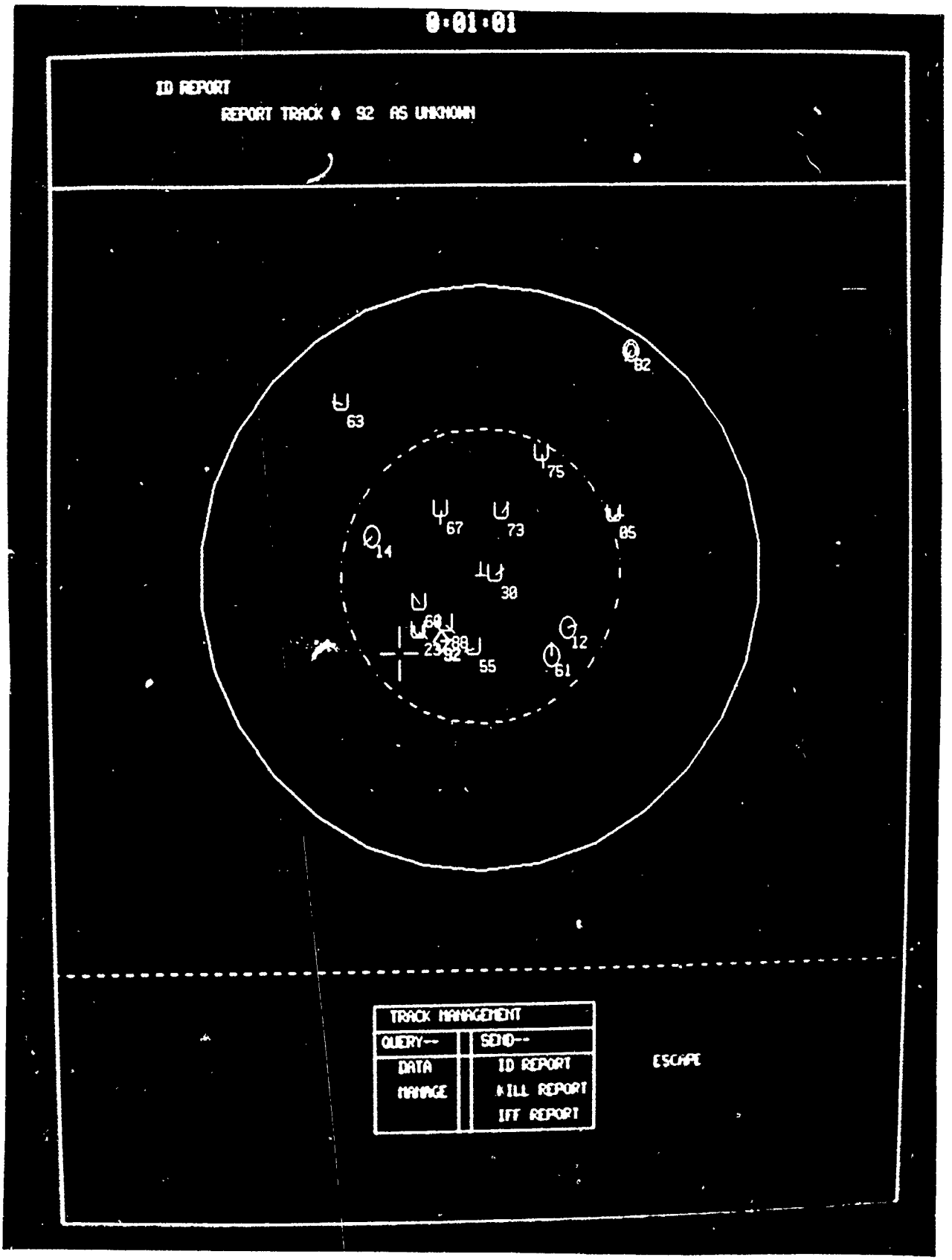

Figure 3. Main menu.

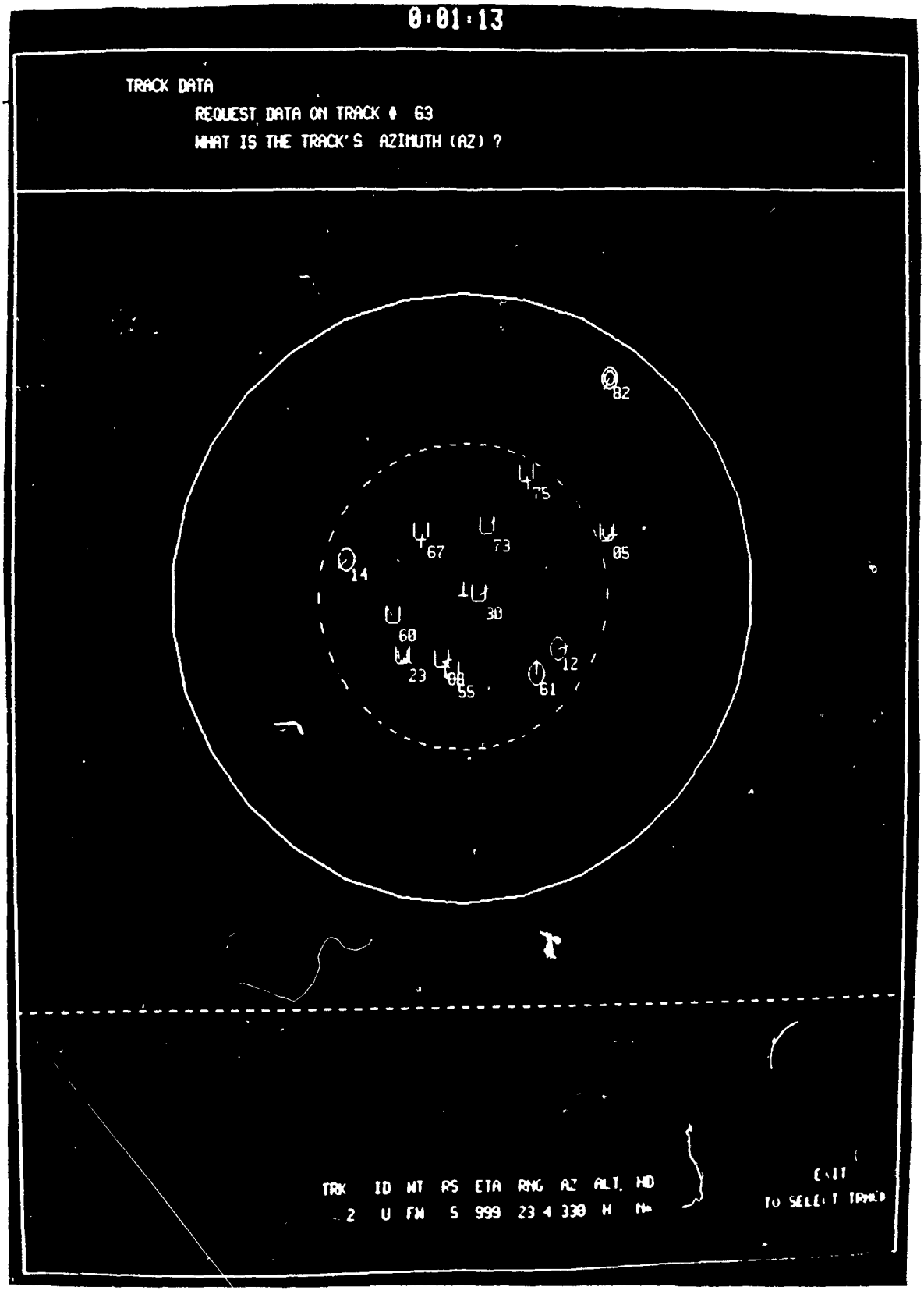

Figure 4. Data sub-menu.

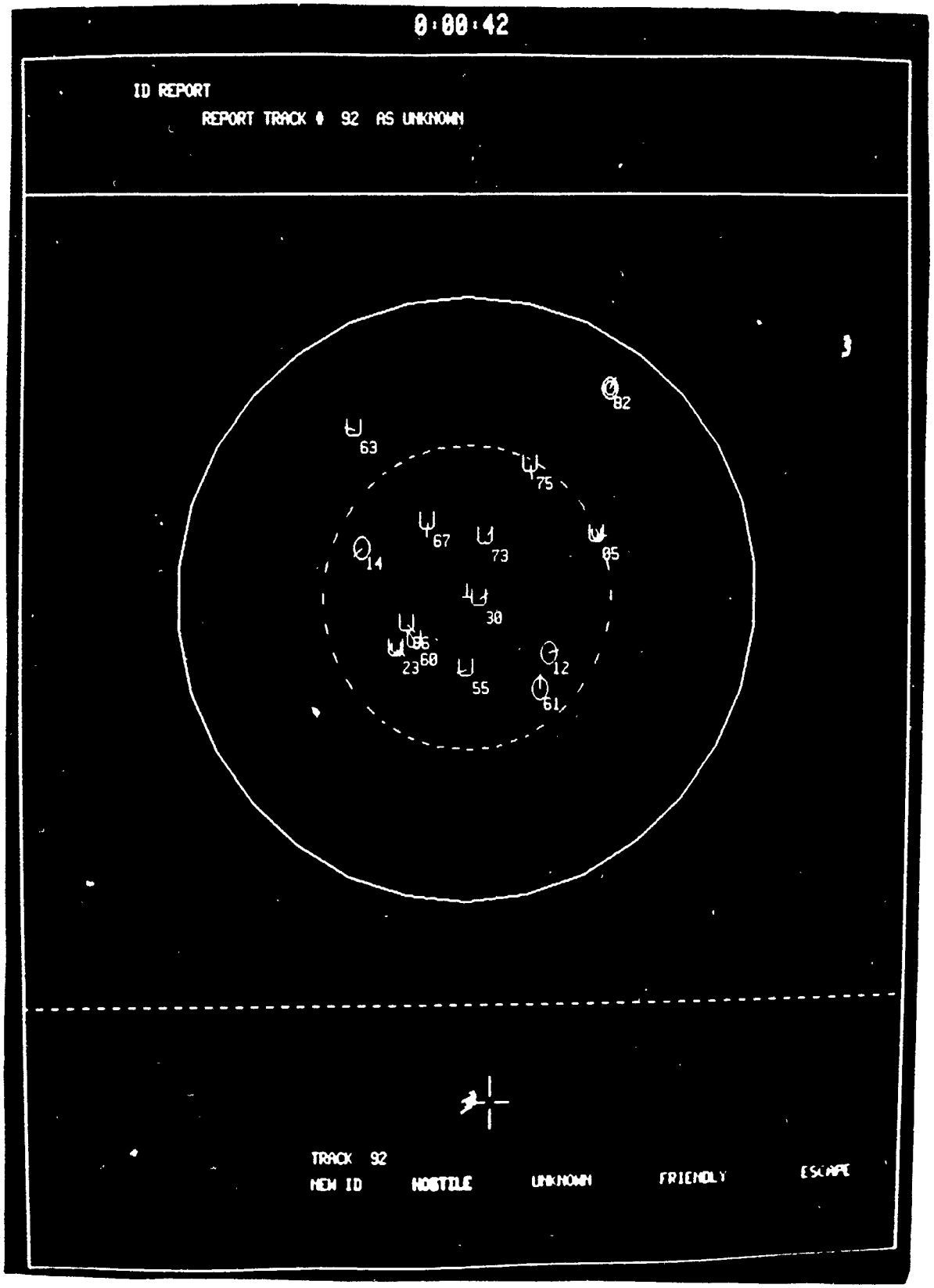

Figure 5. Identification report sub-menu.

The remaining 14 tracks were selected to simulate live aircraft traffic from a randomly chosen set of parameters in accordance with probabilities listed in Table 2, derived from a study of air defense tactical scenarios (Fallesen, Smyth, & Blackmer, 1983). All flight directions were randomly assigned. The initial positions of the tracks in the scenario were randomly selected with the restriction that no positions be closer than 0.5 inch on the display.

Table 2

Scenario Track Probabilities

Factor	Probabilities
Identification	5% hostile 35% friendly 60% unknown
Wing type	35% fixed wing 65% rotary wing
Fixed wing velocities	10% 250 knots 80% 450 knots 10% 450 to 600 knots
altitude	15% low 60% medium 25% high
Rotary wing altitude	100% low (below 0.5 kilometer)
Raid sizes	75% single 25% multiple

The identity of the track was indicated by the symbol shape in accordance with DOD-STD-1477 (DoD, 1983): circular shape for friendly aircraft, diamond shape for hostile aircraft, and U-shape for unknown aircraft. Multiple tracks were shown as two symbols, one inside the other. A line above the symbol indicated a rotary wing track; otherwise, the track was a fixed wing aircraft. The track velocity and direction was shown by a track velocity vector. The track designation number (01 through 99) was displayed to the lower right of the symbol.

The menu display area is at the bottom of the screen, below the graphics display area. This area served as the work area for the subject's interaction with the display. A hierarchical menu method of display interaction was chosen (Miller, 1981). The subject uses a main menu to select sub-menus from which to work. When the subject is finished with a specific task, the subject returns to the main menu to repeat the process for the next task. The advantage of using a hierarchical menu approach is that it provides a logical progression of activity while maintaining the task structure. The potential for user disorganization is avoided by using a low number of menu levels (Billingsley, 1982).

The menu area showed the status of the subject's interaction with the display and listed the menu choices available. The subject's task was limited to information queries about tracks or specification of track identification. In general, the soldier's interaction with the display would be more extensive, encompassing communications about air battle management, battlefield geometry, and command and control messages. For the purposes of this study, the assumption is made that the soldier had selected a "track data management" option from a master menu. The next logical step would be the choice of the specific track for action, followed by a main menu for the action choice and a sub-menu for the action.

The sequence of menu activations on the display for each task was as follows. (Figures 2 through 5 illustrate the changes in the menu area.)

1. The menu area displays a "track select" prompt (see Figure 2). The subject selects the track of interest by voice command (see Table 3). The completion of a track selection action causes the captured track symbol to blink at a 3-hertz rate and the prompt message to be replaced by the main menu.

Table 3

Voice Commands

Data entry stage	Command
Track selection	Hook track XX
Data selection	Hook data
Identification	Hook ID report
Identification selection	Hook friendly Hook unknown Hook hostile
Escape	Escape
Exit from data	Exit

Note. XX represents a two-digit number.

2. The main menu lists the two track action options as sub-menu choices (see Figure 3). The two options open to the subject are "data" for track data amplification and "ID report" for a change in the track's identification status. The completion of a track action selection causes the main menu to be replaced with the corresponding sub-menu.

3. The sub-menu for the track amplification data lists the track number, identification (hostile, unknown, or friendly), wing type (fixed or rotary wing), raid size (single or multiple), estimated time to arrive at host aircraft, range from the host aircraft, azimuth from the host aircraft, altitude (low, medium, or high), and heading (one of the eight cardinal directions) (see Figure 4). Exiting from the sub-menu returns the display to the beginning for a new task.

4. The sub-menu for the track identification task lists the three identification possibilities: friendly, unknown, or hostile (see Figure 5). The current identification of the track is highlighted and selection of the new identification returns the display to the beginning for a new task.

6. In all cases, the option exists for the subject to escape from the main menu or the sub-menu to the "track select" point by using the "escape" voice command. The escape option can be used when the subject selects the wrong track for action, the wrong sub-menu, or becomes confused and wishes to start the process again. Additionally, all menus and sub-menus have cue lines to guide the subject in the available options.

Table 3 lists the voice commands used during each stage of the data selection in the data entry. Table 4 lists the Verbex grammar structure used to process the voice commands. The estimated complexity of the voice commands is 22% of the maximum level as determined by the Verbex voice planner routine (Verbex, 1983). Verbex engineers informed the experimenter that this level is low enough for satisfactory performance.

Table 4

Grammar Definition Table for the Verbex 4000 Voice Planner

Verbex 4000 task grammar
Hook track .digit .digit
Hook .task
:Escape
.Task =
data
id-report
friendly
unknown
hostile
exit
.Digit =
zero
one
two
three
four
five
six
seven
eight
niner

Note. Estimated complexity 22%.

As shown in Table 3, the track symbol was selected from the display area with the voice data entry command "hook track" digit-digit, in which the two digits are the track flight number. The sub-menus were selected from the main menu with the command "hook ID report" for the track identification task, and the command "hook data" for the track data query. The track identification task was completed with one of the commands needed to

change the track identification "hook friendly," "hook unknown," or "hook hostile." The track data query task was completed with the command "exit" after telling the tester the track's flight data. The track data query response was not processed by the automatic speech recognizer. The voice command "escape" was used to clear all entries and return to the start of the task with the track-select menu to correct an error or as an aid in case the subject became confused. All command formats appropriate to the data entry stage were displayed on the screen below the menu location as an aid to the subject.

Test Subjects

The test subjects were 20 military enlisted male personnel of rank E-2 through E-6. The demographics data collected about the subjects are listed in Table 5. All subjects were assigned to the U.S. Army Combat Systems Test Activity (USACSTA), Field Support Branch, Aberdeen Proving Ground, Maryland, and had previous experience as test subjects on military systems. All subjects had little or no experience with computers or speech recognition systems. Four of the subjects had been test subjects during the 1987 study (Smyth et al.).

Experimental Design

A repeated measures factorial experiment with subjects as a random variable was conducted for the training, enrollment, retention, and display variability tests. The test day is the independent variable for the training, enrollment, and retention tests. The independent variable for the variability test is the display update rate which is presented in a random and counterbalanced order. The dependent variable is the number of automatic speech recognizer errors made during the test.

Training and Test Procedures

The subjects were tested 2 hours each day for 5 consecutive days. Each subject was trained and tested individually and all were given the same training. Table 6 shows the training and test schedule for the 5-day session. The training during each of the first 3 days was similar. The subject was first briefed about the test, saw a video tape recording of the task, and enrolled on the automatic speech recognizer. Finally, the subject was trained and tested about the task. On the fourth day, the subject updated his enrollment, practiced the task, and was tested. On the fifth day, the subject, using the previous enrollment, practiced and was tested about the task. The test was then repeated at different display update rates in a counterbalanced order to determine the effects of induced stress because of task variability.

Briefing

The orientation on the first day introduced the subject to the purpose of the test and the test method. He read an explanation of the study and was given the opportunity to ask questions. The subject saw a demonstration film made to show the task being performed and to give the voice commands. A detailed explanation of the tasks and mechanics of the automatic speech recognizer followed.

Table 5

Test Subject Demographic Data

Subject	Rank	MOS[b]	Date of birth (M-D-Y)	Education (years)	Dominant hand	Vision aided
1	E5	88M20	122164	13	Left	No
2	E3	88M10	010565	12	Right	No
3	E4	88M10	092763	12	Left	No
4	E5	19K20	011860	12	Left	contact
5	E3	19E10	072367	12	Left	No
6	E4	19E10	051667	12	Left	No
7	E6	19E30	031058	12	Right	No
8	E4	88M10	070867	12	Right	No
9	E2	88M10	062469	12	Right	No
10	E2	88M10	021467	12	Right	No
11	E2	88M10	101269	12	Right	No
12[a]	E5	11M20	011059	13	Right	No
13	E4	11B10	081165	12	Right	contacts
14[a]	E5	11M20	103163	12	Right	No
15[a]	E5	11B20	073064	13	Right	No
16	E3	13B10	112765	12	Right	No
17	E3	13B10	091367	12	Right	glasses
18[a]	E4	11M10	102164	12	Right	No
19	E4	13B10	060967	12	Right	No
20	E4	88M10	112968	12	Right	No

Note. All males had 20/20 vision or corrected 20/20 vision.

[a] - Prior experience with the automatic speech recognizer in the HEL 1987 test.

[b] - MOS (military occupational speciality) as follows:
11B - Infantryman
11M - Fighting vehicle infantryman
13B - Field artillery cannon crewman
19E - M60 tank armor crewman
19K - M1 tank armor crewman
88M - Motor transport wheeled vehicle operator.

Task level
10 - operations
20 - supervisory
30 - supervisory/management.

Table 6

Training and Test Schedule

Activity	Day 1	Day 2	Day 3	Day 4	Day 5
Brief[a]	O	R	R		
Video[a]	O	R	R		
Enrollment[b]	C	C	C	U	
Train task[c]	T	T	T	P	P
Test task[d]	S	S	S	E	R
Variability task					X

Note. [a] - Brief/video: O - Orientation, R - Review
[b] - Enrollment: C - Complete, U - Update
[c] - Train task: T - Train, P - Practice
[d] - Test task: S - Subject training
E - Enrollment procedure
R - Subject retention

Briefings on the second and third days reviewed the subject's test performance of the previous day. A video tape recording of the subject performing the test was reviewed with the experimenter. The subject was then encouraged to compare the accepted voice commands to those that were rejected by the automatic speech recognizer. The expectancy was that once the subject heard the difference, he would improve the recognizer performance by learning to better control his voice patterns.

Enrollment

The subject established templates of his voice commands in the automatic speech recognizer during the enrollment session before training and testing. The subject was enrolled during each of the first 3 days using the Verbex enrollment training procedures which are machine controlled (Verbex, 1983). The procedures establish a list of word and phrase prompts for the subject to speak. This procedure is from manufacturer-developed programs stored in the machine and the grammar definition table for the task (see Table 4). The Verbex enrollment procedure develops models of the command words and phrases for the subject from his verbal utterances in reply to the machine's prompts. The training procedure refines the word models to accurately represent the continuous speech used by the subject when uttering the command phrases when prompted. The enrollment training procedure took approximately 1 hour to complete using the command phrases of this test. The subject's enrollment was updated on the fourth day (using the Verbex training procedure) based on the word models of the third day. The training procedure took about 30 minutes to complete.

At the start of the enrollment session, the subject was reminded to speak the command phrases in the same way during the enrollment, the task training, and the test sessions. He was advised to relax and was instructed to try for good diction, speaking naturally and forcefully, and pronouncing

each word clearly and distinctly. His speech should flow together naturally; he should avoid long pauses between words. Phrases with long pauses are processed by the automatic speech recognizer as separate unrecognizable phrases instead of the intended command.

The enrollment sessions were rigidly supervised to ensure that the subject's utterances matched the automatic speech recognizer's prompts. The utterances were verified against a checklist constructed before the test from a review of the machine's enrollment prompts. The subject performed his own enrollment following instruction in using VT100 terminal keys to control the enrollment procedure. At the end of the enrollment session, the subject was tested using a series of sample command phrases. The subject was retrained on words or phrases that resulted in less than 100% recognition accuracy.

Control of Extraneous Variables

The subject was advised during the briefing to get enough sleep the night before each test session, to refrain from physical training before each test session, and to refrain from smoking cigarettes and drinking coffee or soft drinks just before the test.

The enrollment training and test were conducted in the same isolated room. Activity in adjacent rooms and hallways was rigidly controlled to prevent extraneous noise. Speech discipline was maintained during the test with speech limited to the task commands and responses by the subject.

A head support system, consisting of a headset and microphone boom holder, was used by the subject to hold the microphone in position. The subject was asked to place the microphone and headset directly on his head in a position that he could duplicate during the following sessions. The microphone should be placed to the side of the mouth and about a thumb width away from the lips. The position of the boom in the holder was measured for reference.

Training

The subject was trained on the task during the first 3 days and practiced the task on the fourth and fifth days before testing. The subject was trained about the interactive, real time computer-driven generic tactical air combat display described previously (Apparatus: Display concepts). This display was used successfully during earlier tests of voice and other data entry methods (Smyth et al., 1987; Dotson & Smyth, 1987). The task times to enter and extract data from the display have proved to be relatively consistent. The subject conducted ten training runs during both training and practice; this number proved appropriate for training during previous studies.

Figure 6 shows a subject wearing the noise-cancelling microphone while sitting at the experimental console in front of the display monitor. The subject used the automatic speech recognizer to give voice commands to the computer to perform data entry and extraction tasks determining the status of track symbology on the display. While performing a task, the subject had to select a track symbol from the tactical display and then select a command line from the menu area to determine data about a track or change its identity.

Figure 6. Test subject control panel.

Each of the training runs was composed of a data entry and a data extraction task in sequence. An instruction line was shown at the top of the display at the start of each task. The prompt instructed the subject to either change the identification of a track on the display or to query the data base for the value of a specified flight parameter: range, altitude, or speed. Using the automatic speech recognizer, the subject selected the track from the tactical display by voice command. He then selected from the main menu the appropriate sub-menu, either track identity or track data, by voice command. The main menu was replaced by the corresponding sub-menu. In one case, the subject specified the track identification update by speech command. This returned the display to the start of the next task. In the second case, the subject told the experimenter the appropriate data item and then voiced the command to reach the next task. See Figure 7 for a schematic diagram of the task flow. The two tasks were assigned to each of the runs in a random order. The track numbers, track data, and identifications were assigned to the tasks in a random and counterbalanced order. All subjects conducted all training runs in a random and counterbalanced order.

Subject Instructions

The instructions to the subject were

> You will see ten training runs and five test runs in each session. Each of the training or test runs is composed of two data entry tasks in succession. An instruction line is shown on the display at the start of each task. The prompt will instruct you to either change the identification of a track or the display or to query the data base for the value of a specified flight parameter (i.e., range, altitude, speed, etc.). You will first select the track from the tactical display by the voice command "hook track" followed by the two digit track number. You will then select the appropriate subtask from a main menu by voice command. That is, you will say "hook ID report" to select data entry for a track identification or "hook data" for data extraction. In either case, the appropriate sub-menu is shown. In the first case, you will select the identification update by voice ("hook friendly," "hook hostile," or "hook unknown"). This will change the track identification and return the display to the start of the next test run. In the second case, you will tell the experimenter the appropriate data item, and then say "exit" to reach the next test run. You may say "escape" at any time to return to the start of the task if you found that you made an error.

Testing

The subject next performed the test on the tactical display. He performed five test runs similar in nature to the training runs. As with training, the two tasks were assigned to each of the test runs in a random order. The track numbers, track data, and identifications were assigned to the tasks in a random and counterbalanced order. All subjects conducted all test runs in a random and counterbalanced order (see Table 7 for the subject test run assignment). The automatic speech recognizer performance was recorded by the experimenter as the subject performed his voice entry tasks. The experimenter aligned and turned on the VCR to record the subject's actions and utterances during the test.

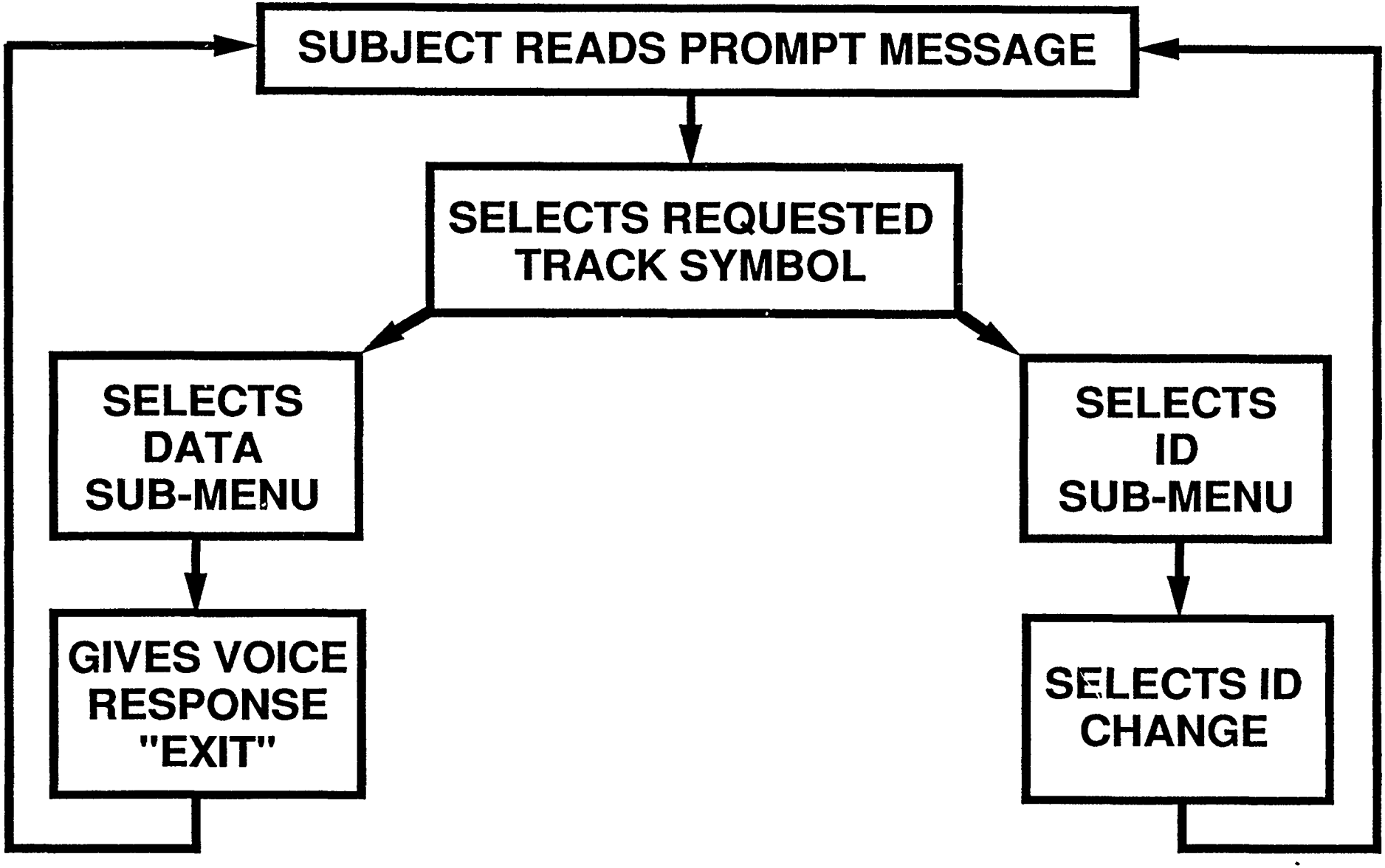

Figure 7. Test subject task flow chart.

Table 7

Testing Sequence Assignments

	Test Trials				
Subject	Day 1	Day 2	Day 3	Day 4	Day 5
1	1	2	3	4	5
2	4	1	5	3	2
3	2	4	1	5	3
4	3	5	2	1	4
5	5	3	4	2	1
6	1	2	3	4	5
7	4	1	5	3	2
8	2	4	1	5	3
9	3	5	2	1	4
10	5	3	4	2	1
11	1	2	3	4	5
12	4	1	5	3	2
13	2	4	1	5	3
14	3	5	2	1	4
15	5	3	4	2	1
16	1	2	3	4	5
17	4	1	5	3	2
18	2	4	1	5	3
19	3	5	2	1	4
20	5	3	4	2	1

Note. Presentation scheme composed of fourfold repetition for five subjects.

On the fifth day, the test was followed by a display update rate test consisting of an additional 15 test runs. This test was similar to previous tests except that the display update rate was accelerated. There were three speeds: normal (1 second update), twice as fast (1/2 second update), and four times as fast (1/4 second update). The tactical display of each test run was updated at one of the display rates, and five test runs were conducted at each update rate. The test runs for the three different update rates were assigned to the subject in a random and counterbalanced order. The subject was told at the start of the test that he would see changes in the display update rate. He was instructed not to accelerate his speech delivery to match the display, but to keep the same voice patterns to ensure that the automatic speech recognizer would continue to work.

STATISTICAL RESULTS

Separate analyses were conducted for the four phases of the test sequence: (1) training phase (Days 1 through 3), (2) enrollment comparison (Days 3 and 4), (3) speech-retention phase (Days 4 and 5), and (4) task variability test (Day 5). The analysis for each of these phases is discussed separately. However, as mentioned previously, the performance of the subjects can be expected to differ (see the "Introduction" section of this report). Therefore, the performance of the subjects, along with the summary statistics of the data, are first reviewed to confirm lack of population homogeneity.

Summary Statistics

The summary statistics for the number of errors for each test period are listed in Table 8. The average number of errors (averaged across subjects) suggest an improvement in performance during the training phase (Days 1 through 3), no change during the enrollment test phase (Days 3 and 4), a decrease during the speech-retention phase (Days 4 and 5), and an increase with increasing display update rate (Rates 1 through 4 during Day 5).

The standard deviations closely track the averages. The standard deviations show wide variation among subjects on Day 1; however, the variation decreases as the average performance increases during the training phase. The variation is contained during the enrollment test phase but increases during the speech-retention phase. Finally, the variation appears to be consistent during the display update rate test. The variances for the different test periods are not homogeneous as shown by Hartley's test (Winer, 1971). In fact, as would be expected for learning data, the variances and averages are strongly correlated with 89.96% of the variance in the averages being explained by the error variances.

Finally, as would be expected for count data, the numbers of errors for each test period are not normal distributions as shown by the skewness and kurtosis parameters (SAS® means procedure) and by the statistical measures of normality (SAS normality procedure).

Subject Grouping

The subjects are not from a homogeneous population but instead form separate clusters according to the number of automatic speech recognizer errors for the test sessions of the training phase. A chi-square analysis (see Table 9) of the total number of errors shows that there is a significant difference among subjects.

Table 9 also shows the subjects grouped into low, medium, and high error subgroups as determined by a chi-square analysis. The subjects were ranked by errors, and the low and medium groupings were based on a sequential nonsignificance chi-square statistic of 11.24 and 9.76, respectively. The remaining subjects were considered to form a high error group. Group I is composed of the seven subjects with the lower errors. Another group (Group II) is composed of six subjects with the next higher number of errors. Finally, the remaining group (Group III) is composed of seven subjects with the highest number of total errors. The group membership is listed in Table 9.

Table 8

Summary Statistics for the Number of Errors for Each Test Period

A. Number of samples, average, standard deviation, standard error skewness, and kurtosis

	Parameters					
Test period	Number	Average	Standard deviation	Standard error	Skewness	Kurtosis
Day 1	20	31.10	51.69	11.56	2.404	6.219
Day 2	20	12.65	27.67	6.19	2.839	6.910
Day 3	20	4.45	4.77	1.07	1.747	3.137
Day 4	20	4.60	7.29	1.63	3.047	11.039
Day 5	20	11.75	12.39	2.77	0.800	-1.060
RateX1	20	12.90	19.37	4.33	2.551	6.551
RateX2	20	13.40	21.42	4.79	2.836	8.504
RateX4	20	14.15	21.09	4.72	2.014	3.406

B. Normality of distribution

	Test Period							
Parameter	Day 1	Day 2	Day 3	Day 4	Day 5	Rate 1	Rate 2	Rate 4
W-Statistic	.658	.449	.799	.630	.795	.609	.604	.670
P < W	.010	.010	.010	.010	.010	.010	.010	.010

C. Correlation of variances to means and homogeneity of variance by the Hartley test

Correlation		
R-square	Pr > T	Hartley's Test
0.8996	0.0003	117.10

Note. Hartley's test statistic, Fmax(.05: 8,20) = 4.10.

Table 9

Chi-Square Analysis of Subject's Total Number of Errors Collapsed Across Training Periods

	Total Number of Errors		
Subject	Observed	Expected	Group
2	304	52.65	III
6	118	52.65	III
19	105	52.65	III
11	102	52.65	III
4	91	52.65	III
5	62	52.65	III
3	48	52.65	III
18	38	52.65	II
1	30	52.65	II
20	27	52.65	II
12	26	52.65	II
9	19	52.65	II
15	19	52.65	II
13	15	52.65	I
10	13	52.65	I
7	12	52.65	I
16	8	52.65	I
17	6	52.65	I
8	5	52.65	I
14	5	52.65	I
Total	1053	1053	

Note. Test value = 1745.8, chi-square statistic (0.001:19) = 36.2.

Some of the error counts are so large that they may be considered extreme outliers based on the nonparametric exploratory data analysis (EDA) technique (Velleman & Hoaglin, 1981). Table 10 shows that these extreme values are concentrated in the high error subject group (Group III) during the training phase (Days 1 through 3), as would be expected since the groups were clustered using the data from this phase. The extreme values do not occur during the enrollment test phase (Day 4) and are scattered among the subject groups during the speech-retention phase (Day 5) and the display update rate test (Day 5). Figure 8 shows the median number of errors for each test session by groups.

Table 10

Exploratory Data Analysis Statistics
(The Number of Extreme Points for Each Subject Group)

Subject group	Number of Extremes by Test Period							
	Day 1	Day 2	Day 3	Day 4	Day 5	Rate 1	Rate 2	Rate 4
I	0	0	0	0	0	0	0	1
II	0	0	0	0	1	2	1	1
III	5	2	0	0	0	1	1	1
Total	5	2	0	0	1	3	2	3

Training Phase

The three subject groups performed differently during the training phase. Group I, with seven subjects, made ten or fewer errors during the test on any day of the training phase and an average of 2.3 errors per day. Most of these errors were spread evenly among the test days. Group II, with six subjects, made 33 or fewer errors during any day and an average of 6.6 errors per day. While most of the errors were spread evenly among the test days, some subjects made a larger number during Days 1 and 2. Finally, Group III, with seven subjects, made an average of 29.6 errors per day. During testing, most subjects made a very large number of errors on Days 1 and 2 of the training phase.

For this reason, the nonparametric Friedman analysis of variance (ANOVA) by ranks for repeated measures was applied separately to the data of the three subject groups. The results of the separate analyses are listed in Table 11. The analysis is extended to include the data of Day 4 for reasons explained in the following Enrollment Phase section. The variation in the number of recognition errors for Group III is statistically significant (0.02 level). A nonparametric post hoc contrast test (Marascuilo & McSweeney, 1977) shows that the number of errors for Day 1 is significantly greater (0.05 level) than those for the remaining days. In contrast, the differences among the number of errors for the training days are insignificant for Groups I and II.

Apparently, at least two training sessions were needed by seven of the subjects (35%) before they could perform the task at their best ability. However, the additional training did not significantly improve the remaining subjects' performances.

Enrollment Phase

While the full enrollment procedure was used on Day 3, the subject used only the recognizer-training portion of the enrollment on Day 4. The number of errors generated during these two days are not significantly different (0.05 level) according to a parametric ANOVA. The assumptions for a parametric study are satisfied; the variances are homogeneous and not correlated with the means, and the distributions are normal. See Table 12 for summary statistics. The correlation for an accompanying parametric regression

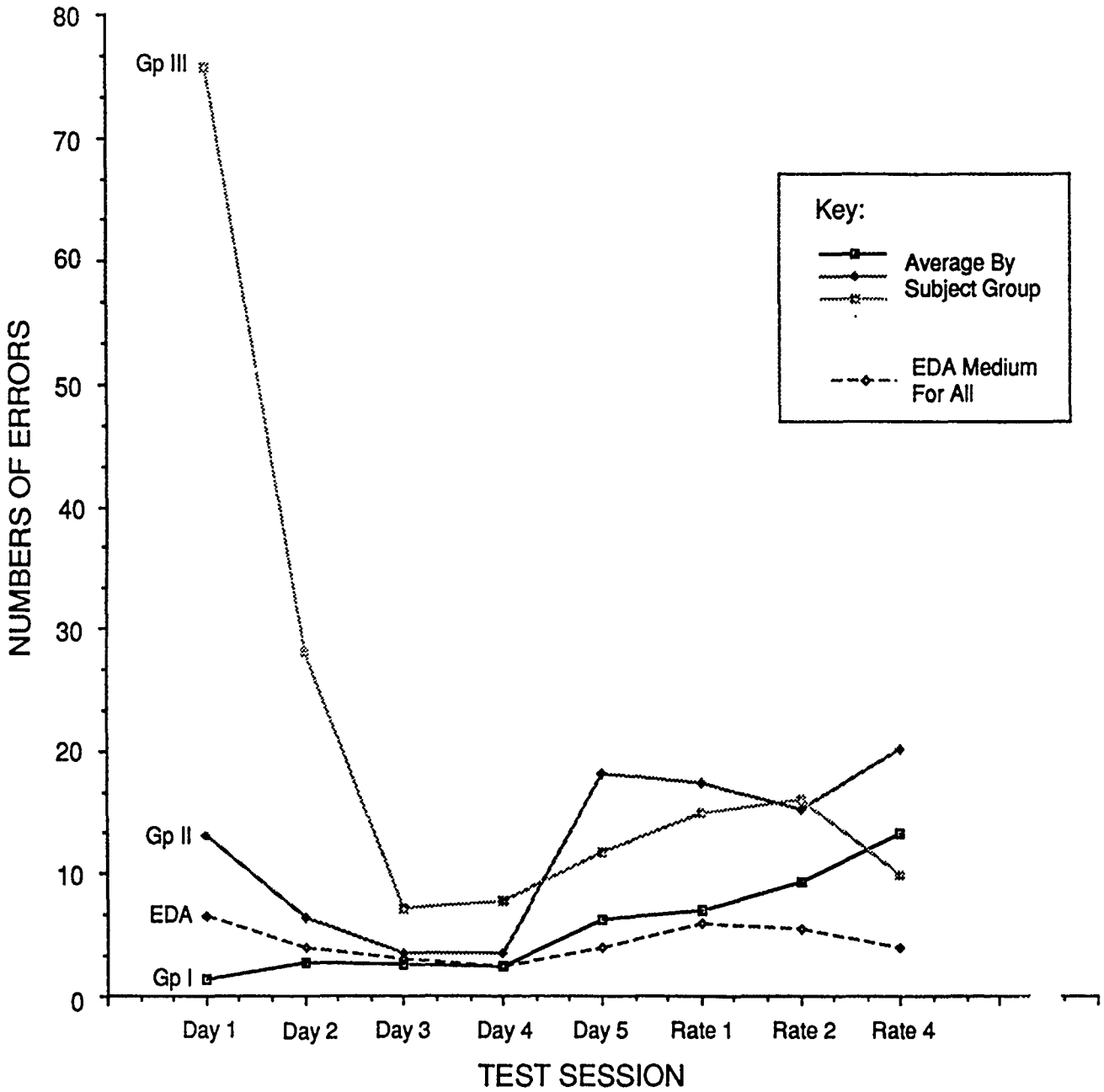

Figure 8. Number of errors by test session.

Table 11

Friedman's Nonparametric Two-Way Analysis of Variance by Ranks for the Number of Errors for Each of the Three Subject Groups for the Training Periods

	Number of Errors				Ranking			
Subject	Day 1	Day 2	Day 3	Day 4	Day 1	Day 2	Day 3	Day 4
				Group I				
13	2	4	6	3	4	2	1	3
10	0	2	1	10	4	2	3	1
7	4	2	6	0	2	3	1	4
16	0	8	0	0	3	1	3	3
17	1	0	2	3	3	4	2	1
8	2	2	0	1	1.5	1.5	4	3
14	1	1	3	0	2.5	2.5	1	4
Total	10	19	18	17	20	16	15	19
				Group II				
18	33	4	1	0	1	2	3	4
1	18	4	5	3	1	3	2	4
20	20	7	0	0	1	2	3	4
12	1	12	3	10	4	1	3	2
9	4	4	5	6	3.5	3.5	2	1
15	3	7	7	2	3	1.5	1.5	4
Total	79	38	21	21	13.5	13	14.5	19

Table 11 (continued)

	Number of Errors				Ranking			
Subject	Day 1	Day 2	Day 3	Day 4	Day 1	Day 2	Day 3	Day 4
				Group III				
2	204	97	1	2	1	2	4	3
6	117	0	1	0	1	3.5	2	3.5
19	9	87	7	0	2	1	3	4
11	77	2	15	8	1	4	2	3
4	66	1	18	6	1	4	2	3
5	48	5	3	6	1	3	4	2
3	9	2	5	32	2	4	3	1
Total	530	194	50	54	9	21.5	20	19.5

Note. Group I: Friedman test value = 1.46
Chi-square statistic (0.05: 3) = 7.81

Group II: Friedman test value = 2.25
Chi-square statistic (0.05: 3) = 7.81

Group III: Friedman test value = 8.44
Chi-square statistic (0.05: 3) = 7.81

Post Hoc Contrast Tests
Coefficients

Contrasts	Day 1	Day 2	Day 3	Day 4	Value	Variance	Range
linear	-3	-1	+1	+3	4.28	4.76	-1.81,+10.37
quadratic	+1	-1	-1	+1	-1.86	0.95	-4.58,+ 0.87
Day 1 vs Day 2-4	-3	+1	+1	+1	4.86	2.86	+0.13,+ 9.58
Days 1&2 vs Days3&4	-1	-1	+1	+1	1.28	0.95	-1.44,+ 4.01

Table 12

Table of Summary Statistics for the Analysis of Variance for the Enrollment Test Phase

Table of Summary Statistics

Source	DF	SS	MS	F	Pr >F	R-square
Model	1	0.225	0.225	0.01	0.939	0.0001
Error	38	1441.75	37.94			
Total	39	1441.97				

analysis is nonsignificant. Apparently, the use of enrollment training in place of full enrollment, following proper establishment of speaker templates, did not affect performance. For this reason, the training phase data were considered for statistical analysis to be from Days 1 through 4 and the speech-retention phase from Days 3 through 5 to increase the degrees of freedom and therefore, the power of the analyses.

Speech-Retention Phase

The number of errors across all subject groups combined during the speech-retention phase is significantly different (0.05 level) between days. This is according to a nonparametric Friedman ANOVA by ranks for repeated measures. A post hoc contrast test applied to the data shows that the number of errors for Day 5 is significantly greater (0.05 level) than those for the preceding two days (see Table 13). The analysis is extended to include the data from Day 3 for reasons previously given in the Enrollment Phase section. The results show that the performance of subjects was significantly reduced when they used the enrollment from the previous day.

Task Variability Test

The number of errors for the task variability test, summed across all subjects, shows a nonsignificant increase with an increase in the display update rate as determined by a nonparametric Friedman ANOVA by ranks. The nonsignificance results from the large variation in data across all three groups (see Table 14).

Interestingly, the nonparametric Friedman analysis of the number of errors for the Day 5 speech-retention phase and task variability tests also shows nonsignificant differences. However, the variance for the Day 5 speech-retention phase is significantly smaller (0.05 level) than that for the task variability tests pooled together. The Hartley test (Winer, 1971) shows that while the variances for the task variability tests are homogeneous, the variances for the Day 5 speech-retention phase and pooled task variability tests are not (see Table 15).

In summary, changes in a task parameter, such as the display update rate, significantly increases the variation in recognizer performance.

Table 13

Friedman's Nonparametric Two-Way Analysis of Variance by Ranks for the Number of Errors for the Retention Test Periods

	Number of Errors			Ranking		
Subject	Day 3	Day 4	Day 5	Day 3	Day 4	Day 5
1	5	3	37	2	3	1
2	1	2	2	3	1.5	1.5
3	5	32	22	3	1	2
4	18	6	28	2	3	1
5	3	6	4	3	1	2
6	1	0	0	1	2.5	2.5
7	6	0	2	1	3	2
8	0	1	9	3	2	1
9	5	6	30	3	2	1
10	1	10	3	3	1	2
11	15	8	25	2	3	1
12	3	10	6	3	1	2
13	6	3	3	1	2.5	2.5
14	3	0	23	2	3	1
15	7	2	28	2	3	1
16	0	0	1	2.5	2.5	1
17	2	3	3	1	2.5	2.5
18	1	0	4	2	3	1
19	7	0	1	1	3	2
20	0	0	4	2.5	2.5	1
Total	89	92	235	43	46	31

Note. Friedman test value = 6.30, chi-square statistic (0.05: 2) = 5.99.

Post Hoc Contrast Tests
Coefficients

Contrast	Day 3	Day 4	Day 5	Value	Variance	Range
Day 3, Day 4 vs Day 5	-1	-1	+2	-1.35	0.30	-2.69, -0.01

Table 14

Friedman's Nonparametric Two-Way Analysis of Variance by Ranks for the Number of Errors for the Day 5 Task Consistency Test

	Number of Errors			Ranking		
Subject	Rate 1	Rate 2	Rate 4	Rate 1	Rate 2	Rate 4
1	38	52	29	2	1	3
2	3	3	1	3	1	2
3	6	8	19	3	2	1
4	4	1	4	1.5	3	1.5
5	7	6	4	1	2	3
6	6	1	2	1	3	2
7	4	11	2	2	1	3
8	13	14	63	3	2	1
9	46	25	74	2	3	1
10	17	18	20	3	2	1
11	78	89	35	2	1	3
12	6	1	3	1	3	2
13	2	3	1	2	1	3
14	3	5	0	2	1	3
15	1	2	6	3	2	1
16	4	12	3	2	1	3
17	6	2	4	1	3	2
18	7	10	7	2.5	1	2.5
19	4	4	4	2	2	2
20	6	1	2	1	3	2
Total	258	268	283	40	38	42

Note. Friedman test value = 0.21, chi-square statistic (0.05: 2) = 5.99.

Table 15

Hartley's Homogeneity Test Applied to the Day 5 Retention and Pooled Task Consistency Test Periods (Pooled Rate 1, Rate 2, & Rate 4)

Period	Number	Average	Standard deviation	Variance
Day 5	20	11.75	12.39	153.51
Pooled	60	13.48	20.30	412.09

Note. Test value = 2.68, Hartley's Fmax statistic (0.05:20,2) = 2.46.

DISCUSSION

This section of the report reviews (a) the recognition frequencies as a function of test sessions, (b) the automatic speech recognizer errors, (c) the possible effects of the test procedure including practice and task interruptions on performance, (d) the task consistency test, (e) the effects of subject errors on recognizer performance, (f) the relation of military experience and training to performance, and (g) a summary including a few observations and the subjects' ratings of the automatic speech recognizer's usefulness to Army systems.

Recognition Frequency

Figure 9 is a plot of the correct recognition frequency averaged across all subjects for each of the test sessions. The plot shows the session averages and confidence intervals useful for nonparametric statistics. Table 16 lists the corresponding summary statistics for the frequency of recognition. The percentages are computed from the number of correct recognitions for a subject and divided by his total number of command phrases to the automatic speech recognizer during the test session.

The frequency plot agrees with the results section. The average performance of the voice automatic speech recognizer during the first test session was the low value of 70.56%. The confidence interval (0.05 level), containing 95% of the subjects, varies from 57.62% through 83.50%. Apparently, through the adaptation process provided by the training phase, the subjects learned to maintain speech patterns consistently enough to successfully operate the automatic speech recognizer. The frequency plot for the training phase shows a negatively accelerating asymptotic curve which is usually associated with learning.

The average performance after full training was 89.62% recognition accuracy with a 84.31% through 94.93% confidence interval. This is consistent with results of a previous HEL experiment (Malkin & Christ, 1985) and other laboratories with operational environments.

The average performance without daily enrollment was 78.10% with a confidence interval of 69.97% through 86.23%. This was a decrease in performance of 11.52%. The frequency plot for the speech-retention phase shows a classical extinction curve.

Automatic Speech Recognizer Errors

This portion of the report reviews the mis-recognitions for the first five test sessions as follows: (a) frequency by error types, (b) frequency by command phrases, (c) variations with test trial, (d) nature of mis-recognitions, (e) initial mis-recognitions, and (f) sequences of mis-recognitions. Mis-recognitions for the task variability test are discussed in a separate section of this report.

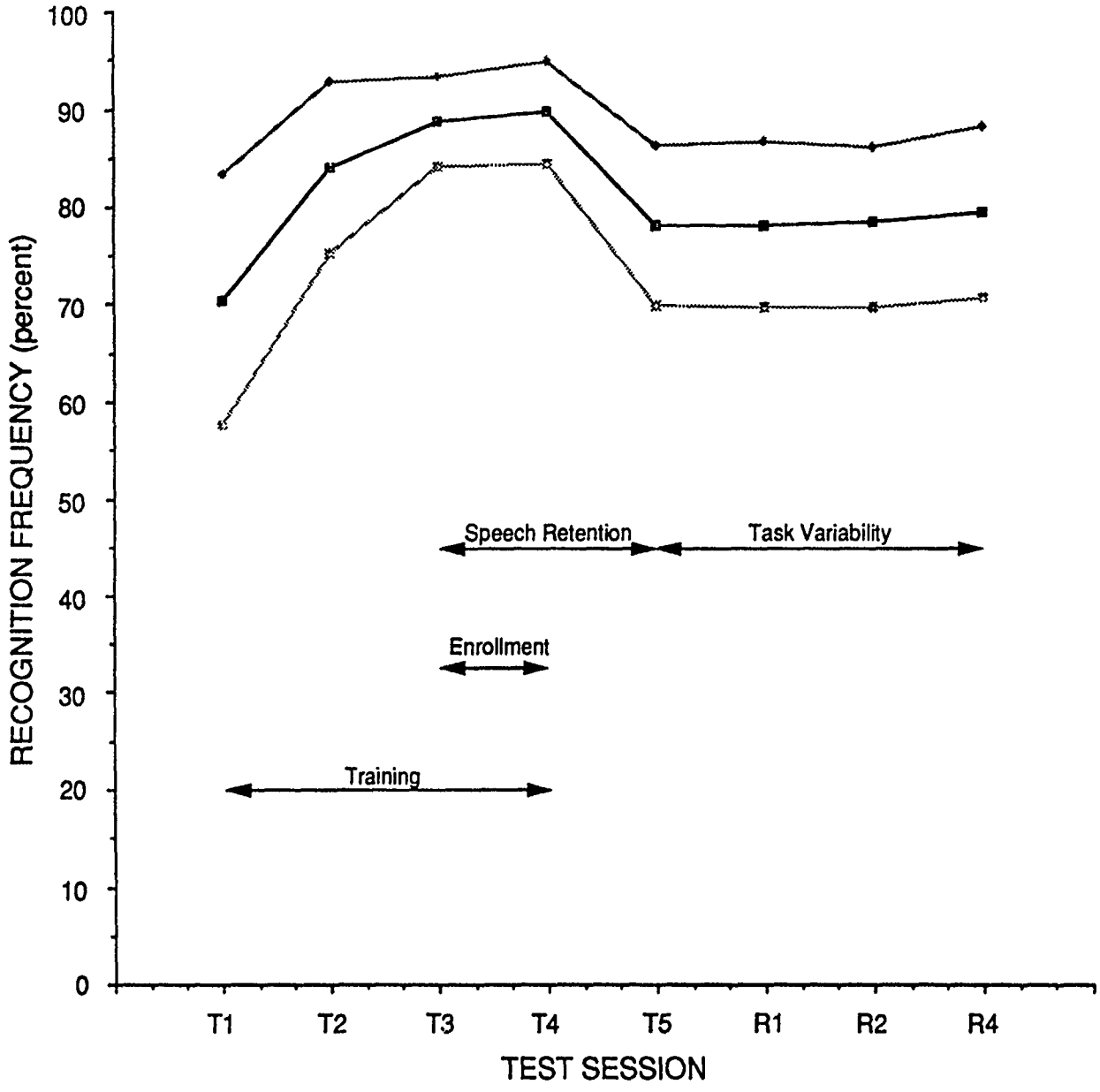

Figure 9. Recognition frequency (percent) by test session.

Table 16

Summary Statistics for the Frequency of Correct Recognition (Percentage) for Each Test Period

	Parameters				Confidence Interval	
Test period	Number	Average	Standard deviation	Standard error	Lower	Upper
Day 1	20	70.56	29.51	6.60	57.62	83.50
Day 2	20	83.98	20.10	4.49	75.18	92.78
Day 3	20	88.79	10.40	2.32	84.24	93.34
Day 4	20	89.62	12.10	2.71	84.31	94.93
Day 5	20	78.10	18.58	4.15	69.97	86.23
RateX1	20	78.17	19.36	4.33	69.68	86.66
RateX2	20	78.40	19.95	4.46	69.66	86.14
RateX4	20	79.43	20.15	4.51	70.59	88.27

Note. Confidence interval 0.05 level.

Error by Type

On the average, 97.99% of the errors were nonrecognitions, only 2.01% were substitutions, and none were rejections or insertions. A nonrecognition error was committed when the automatic speech recognizer could not classify an utterance as a training set phrase. A substitution occurred when the machine incorrectly recognized a phrase. A rejection would be the nonrecognition of a partial phrase, and an insertion would be the recognition of a partial phrase which was not spoken. These statistics are fairly consistent across test sessions as shown in Table 17, listing the percentage of each type of mis-recognition possible (Pallett, 1985) committed by the automatic speech recognizer (given that the subject uttered the correct command phrase). The table lists the percentages for each of the first five test sessions and all these sessions combined.

Table 17

Percentage of Automatic Speech Recognizer Errors by Error Type

	Test Session					
Error type	Day 1	Day 2	Day 3	Day 4	Day 5	Combined
Nonrecognition	99.04	96.03	97.78	98.11	96.93	97.99
Substitution	0.96	3.97	2.22	1.89	3.07	2.01
Rejection	0	0	0	0	0	0
Insertion	0	0	0	0	0	0

Error by Command Phrase

The errors were not caused by any one command phrase, but scattered among them. This is shown in Table 18 which lists the percentage of mis-recognition by command phrase for each of the first five test sessions and for the theoretical distribution assuming that the errors occur randomly for all phrases. The table also lists the results of a chi-square contingency table analysis; a significant interaction exists between the test sessions and the command phrases. A study of the percentages suggests that more than the expected errors occurred for the phrase "hook ID report" during Days 1 and 2, "exit" during Day 3, "hook friendly" during Day 4, and for the phrases "hook track" and "hook data" during Day 5. A study of the data shows that these errors are caused by a few subjects experiencing long sequences of mis-recognitions; otherwise, the errors are evenly spread among the command phrases for all test runs.

Error by Test Run

The errors are distributed randomly among the test runs of each test session; this suggests that performance did not improve during each test session. Table 19 lists the number of automatic speech recognizer errors by test runs for each of the first five test sessions and for the test sessions combined. A nonparametric Friedman ANOVA by ranks shows nonsignificant differences by test runs. Apparently, training occurred before each test and not during the test for that day.

Nature of Mis-recognitions

Most mis-recognitions were followed by a successful recognition during the next utterance. These singular mis-recognitions produced only a few errors per test run. However, consecutive runs of multiple mis-recognitions by the automatic speech recognizer occasionally occurred. These runs of multiple mis-recognitions are the source of extreme numbers of errors in the data.

These occurrences were not because of any one subject; however, the training apparently reduced the occurrence. The multiple mis-recognitions were experienced by the high error subject group during the training test phrase and by all groups during the speech-retention phase. These runs caused the extreme number of errors found by EDA for these two phases. These mis-recognition runs rarely occurred after full training and daily enrollment (Days 3 and 4). The mis-recognition runs appeared to result from changes in the voice patterns from the enrollment patterns, of which the subjects were apparently unaware.

The number of mis-recognitions for each test session as a function of subject and test run is listed in the Appendix. As mentioned previously, the data show that a few mis-recognitions occurred during most test runs; however, occasionally some subjects experienced a large number of mis-recognitions during one test run. An EDA shows that more than eight mis-recognitions during one test run should be considered extreme values. The following sections discuss the statistics for first mis-recognitions and then the number of repeats needed to obtain recognition.

Table 18

Automatic Speech Recognizer Errors by Command Phrase

Percentage of Errors

Command phrase	Day 1	Day 2	Day 3	Day 4	Day 5	Theoretical
Hook track	25.72	38.49	28.09	23.58	53.51	33.33
Hook data	12.54	17.85	8.89	13.21	6.58	16.67
Exit	5.63	6.75	30.34	17.92	11.40	16.67
Hook ID report	37.62	28.17	16.85	14.15	14.91	16.67
Hook friendly	8.84	3.17	8.99	25.47	5.70	5.55
Hook unknown	6.44	1.19	3.37	1.89	3.51	5.55
Hook hostile	2.90	4.37	3.37	3.77	2.63	5.55
Escape	0.32	0.0	0.0	0.0	1.75	0.0

Chi-Square Analysis

Command phrase	Day 1	Day 2	Day 3	Day 4	Day 5	Total
			Observed Contingency Table			
Hook track	160	97	25	25	122	429
Hook data	78	45	8	14	15	160
Exit	35	17	27	19	26	124
Hook ID report	234	71	15	15	34	369
Hook friendly	55	8	8	27	13	111
Hook unknown	40	3	3	2	8	56
Hook hostile	18	11	3	4	6	42
Total	620	252	89	106	224	1291
			Expected Contingency Table			
Hook track	206.6	84.0	29.7	35.3	74.7	430.3
Hook data	103.4	42.0	14.8	17.7	37.3	215.2
Exit	103.4	42.0	14.8	17.7	37.3	215.2
Hook ID report	103.4	42.0	14.8	17.7	37.3	215.2
Hook friendly	34.4	14.0	4.9	5.9	12.4	71.6
Hook unknown	34.4	14.0	4.9	5.9	12.4	71.6
Hook hostile	34.4	14.0	4.9	5.9	12.4	71.6
Total	620	252	88.8	106.1	223.8	1290.7

Note. Test value = 449.20, chi-square statistic (0.05:24) = 36.4.

Table 19

Number of Automatic Speech Recognizer Errors by Test Run

Test session	Run 1	Run 2	Run 3	Run 4	Run 5	Combined
			Test Runs			
			Number of Errors			
Day 1	122	54	198	126	122	622
Day 2	26	71	47	44	64	252
Day 3	24	21	24	12	16	97
Day 4	14	37	17	27	16	111
Day 5	75	30	44	29	57	235
Combined	261	213	330	238	275	1317
			Friedman's Nonparametric Analysis of Variance by Ranks			
Day 1	3.5	5.0	1.0	2.0	3.5	
Day 2	5.0	1.0	3.0	4.0	2.0	
Day 3	1.5	3.0	1.5	5.0	4.0	
Day 4	5.0	1.0	3.0	2.0	4.0	
Day 5	1.0	4.0	3.0	5.0	2.0	
Total	16.0	14.0	11.5	18.0	15.5	

Note. Friedman test value = 1.88, chi-square (0.05:4) = 9.49.

Initial Mis-recognitions

The initial mis-recognitions show a nonsignificant decrease in the number of errors with test session during the training phase and a significant increase during the speech-retention phase. Table 20 lists the summary statistics for the number of first mis-recognitions for the first five test sessions. The table shows that the average number of errors closely track the median values listed in Table 10 for EDA. The standard deviations are close to the corresponding averages. The probability of first recognition is not significantly different from the recognition performance shown in Table 16.

Table 21 lists the results of a nonparametric Friedman two-way ANOVA by ranks for the number of first mis-recognitions for the training phase. The results verify that the number of first mis-recognitions do not significantly vary with test session. Similar results occur for an analysis by subject group for the training phase.

In contrast, Table 22 shows the results of a Friedman test applied to the number of first mis-recognitions of the speech-retention phase (Days 4 and 5). The results show a significant increase in the number of errors without daily enrollment.

Table 20

Summary Statistics for the Number of First Recognition Errors for the Training, Enrollment, and Retention Test Periods

Test period	Number	Average	Standard deviation	Standard error	Probability correct	Probability error
			Parameters			
Day 1	20	5.50	5.41	1.01	81.67	18.03
Day 2	20	4.45	4.90	0.91	85.16	16.30
Day 3	20	3.25	3.42	0.95	89.17	11.40
Day 4	20	2.90	2.90	1.00	90.33	9.67
Day 5	20	5.05	4.32	1.17	83.17	14.40

Mis-recognition Repeats

The number of repeats needed for recognition decreased with each day of the training phase but increased during the speech-retention phase during Day 5. Table 23 lists the number of repeats needed for recognition, given an initial mis-recognition, for each of the first five test sessions. The table shows that recognition occurred during one repeat following 53.6% of the initial mis-recognitions on Day 1, but occurred following 82.8% on Day 4. In contrast, recognition during one repeat drops to 61.4% of the initial mis-recognitions on Day 5. A chi-square test for independent samples shows significant interaction between the numbers of repeats and test sessions. The numbers of repeats have been combined in the contingency table to ensure that none of the the expected values are less than five to satisfy restrictions for analysis (Siegel, 1956).

Figure 10 shows the data of Table 23 in frequency plots for each test period. The figure demonstrates that while the repeats needed for recognition following mis-recognition may be quite large for the first two test periods (Days 1 and 2), the repeats are reduced to practically one or two for the third and fourth test periods. However, the repeats for the fifth test period, when prior enrollment was not used, increased to the level of the initial training periods.

Table 24 shows the distributions of repeats needed by subject groups for each test period. The table shows that while Group III had the most repeats during the training phase, they improved with each session as did the other two groups. However, the repeats increased drastically for the three groups during Day 5. A chi-square analysis shows that significant differences occurred in the repeats by test sessions for Group II.

Table 21

Friedman's Nonparametric Two-Way Analysis of Variance by Ranks for the Number of First Recognition Errors for the Training Periods

	Number of Errors				Ranking			
Subject	Day 1	Day 2	Day 3	Day 4	Day 1	Day 2	Day 3	Day 4
1	4	4	5	3	2.5	2.5	1	4
2	6	21	1	2	2	1	4	3
3	8	2	2	7	1	3.5	3.5	2
4	12	1	13	5	2	4	1	3
5	12	5	3	5	1	2.5	4	2.5
6	19	0	1	0	1	3.5	2	3.5
7	3	2	3	0	1.5	3	1.5	4
8	2	2	0	1	1.5	3	1.5	4
9	3	4	5	6	4	3	2	1
10	0	2	1	7	4	2	3	1
11	15	2	10	7	1	4	2	3
12	1	7	3	8	3	2	3	1
13	2	4	4	3	4	1.5	1.5	3
14	1	1	2	0	2.5	2.5	1	4
15	3	5	3	2	2.5	1	2.5	4
16	0	7	0	0	3	1	3	3
17	1	0	1	2	2.5	4	2.5	1
18	10	4	1	0	1	2	3	4
19	3	13	7	0	3	1	2	4
20	5	3	0	0	1	2	3.5	3.5
Total	110	89	65	58	45	49	47.5	58.5

Note. Friedman test value = 3.14, chi-square statistic (0.05: 3) = 7.81.

Table 22

Friedman's Nonparametric Two-Way Analysis of Variance by Ranks for the Number of First Recognition Errors for the Speech-Retention Periods

	Number of Errors			Ranking		
Subject	Day 3	Day 4	Day 5	Day 3	Day 4	Day 5
1	5	3	15	2	3	1
2	1	2	2	3	1.5	1.5
3	2	7	9	3	2	1
4	13	5	12	1	3	2
5	3	5	4	3	1	2
6	1	0	0	1	2.5	2.5
7	3	0	2	1	3	2
8	0	1	3	3	2	1
9	5	6	13	3	2	1
10	1	7	3	3	1	2
11	10	7	8	1	3	2
12	3	8	4	3	1	2
13	4	3	2	1	2	3
14	2	0	8	2	3	1
15	3	2	5	2	3	1
16	0	0	1	2.5	2.5	1
17	1	2	3	3	2	1
18	1	0	4	2	3	1
19	7	0	1	1	3	2
20	0	0	2	2.5	2.5	1
Total	65	58	101	43	46	31

Note. Friedman test value = 6.30, chi-square statistic (0.05: 2) = 5.99.

Post Hoc Contrast Tests
Coefficients

Contrast	Day 3	Day 4	Day 5	Value	Variance	Range
Day 3, Day 4 vs Day 5	-1	-1	+2	-1.35	0.30	-2.69, -0.01

Table 23

Number of Repeats Needed for Recognition Given Mis-Recognition for Each Test Period

Test period	Number of repeats 1	2	3	4	5	6	>6
				Percentage			
Day 1	53.6	11.8	8.2	5.4	0.9	1.8	18.3
Day 2	73.0	5.6	4.5	1.1	2.2	4.5	9.1
Day 3	75.4	15.4	6.2	3.0			
Day 4	82.8	10.3	5.2	1.7			
Day 5	61.4	14.8	7.9	3.0	4.9	2.0	6.0
				Frequency			
Day 1	59	13	9	6	1	2	20
Day 2	65	5	4	1	2	4	8
Day 3	49	10	4	2			
Day 4	48	6	31				
Day 5	62	15	8	3	5	2	6

Chi-Square Test

Test period	Number of repeats 1	2	3-4	4->6	Total
			Observed Values Combined		
Day 1	59	13	15	23	110
Day 2	65	5	5	14	89
Day 3	49	10	6	0	65
Day 4	48	6	3	1	58
Day 5	62	15	11	13	101
Total	283	49	40	51	423
			Expected Values Combined		
Day 1	75.15	12.74	10.40	13.26	110
Day 2	60.81	10.31	8.42	10.73	89
Day 3	44.41	7.53	6.15	7.84	65
Day 4	39.63	6.72	5.48	6.99	58
Day 5	75.15	11.70	9.55	12.18	101
Total	283	49	40	51	423

Note. Test value = 38.81, chi-square statistic (0.05:12) = 21.0.

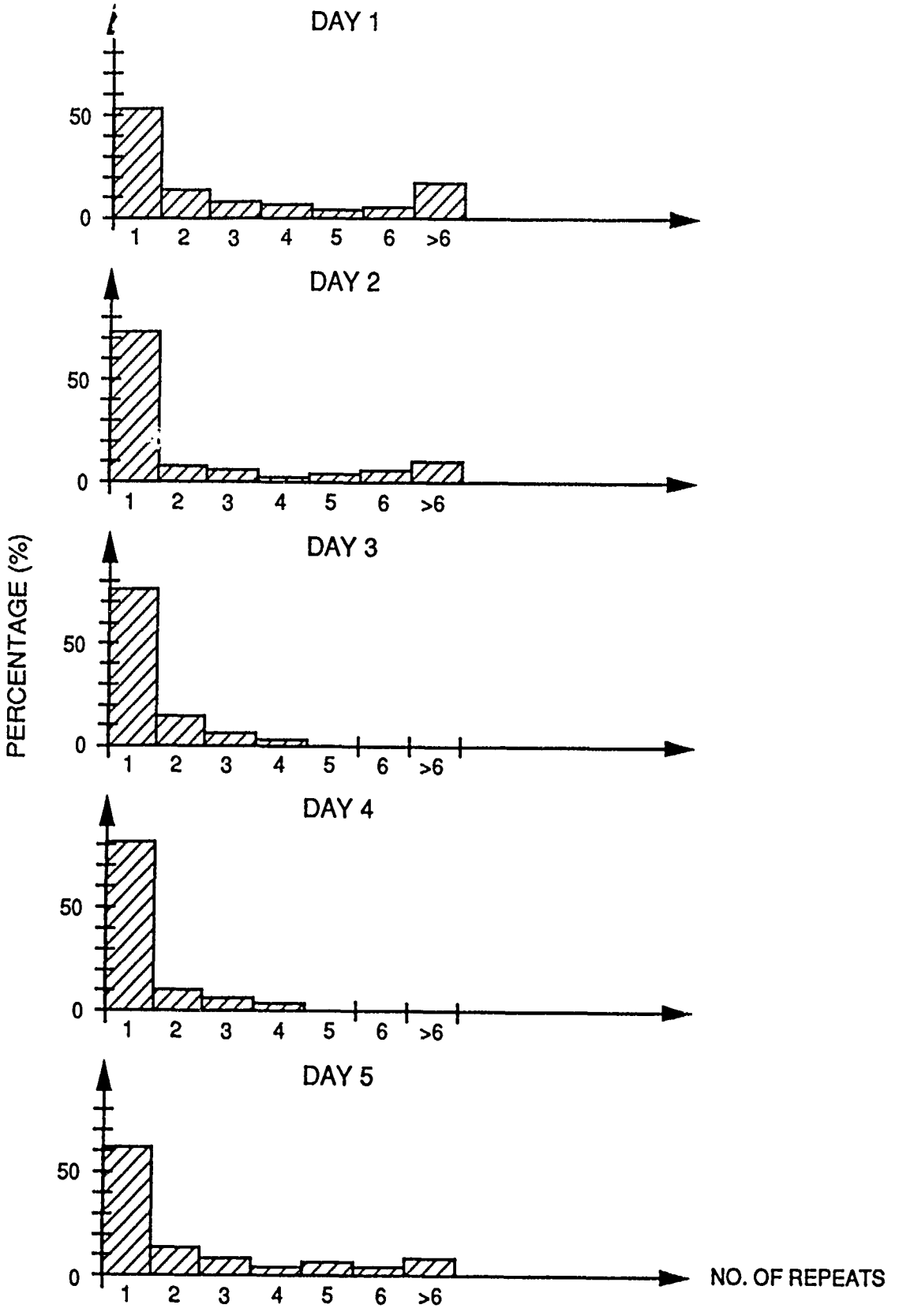

Figure 10. Perentage number of repeats for recognition given mis-recognition for each test period.

Table 24

Frequency of the Number of Repeats Needed for Recognition Given Mis-Recognition for Each Test Period by Subject Groups

Test period	Number of repeats 1	2	3	4	5	6	>6
			Group I				
Day 1	8	1					
Day 2	17	1					
Day 3	7	2	1	1			
Day 4	10	2	1				
Day 5	14	2	3		1	2	
			Group II				
Day 1	15	4	2	2			1
Day 2	22	1	3		1		
Day 3	25	1		1			
Day 4	18	0	1				
Day 5	26	7	3	1	3		1
			Group III				
Day 1	36	8	7	4	1	2	4
Day 2	26	3	1	1	1		
Day 3	27	7	3				
Day 4	20	4	1				
Day 5	22	6	2	2	1		2

Table 24 (Continued)

Chi-Square Analysis

Test period	Group I	Group II			Group III		
		Combined repeats			Combined repeats		
	All	1	2->6	Total	1	2->6	Total
			Observed				
Day 1	9	15	9	24	36	26	62
Day 2	18	22	5	27	26	6	32
Day 3	11	25	2	27	27	10	37
Day 4	13	18	1	19	20	5	25
Day 5	22	26	15	41	22	13	35
Total	73	106	32	138	131	60	191
			Expected				
Day 1	14.6	18.43	5.57	24	42.52	19.48	62
Day 2	14.6	20.74	6.26	27	21.95	10.05	32
Day 3	14.6	20.74	6.26	27	25.38	11.62	37
Day 4	14.6	14.59	4.41	19	17.15	7.85	5
Day 5	14.6	31.49	9.51	41	24.01	10.99	35
Total	73	106	32	138	131	60	191
Test values	7.75		14.41			7.93	

Note. Chi-square statistic (0.05:4) = 9.49.

Test Procedure Effects

The possible effects of the test procedure on the recognizer performance are discussed in this section. These include the results of a test sequence analysis to determine possible changes in test procedure over time, the need for practice during the speech-retention phase of Day 5, and the possible effects of the interruption between practice and test on recognizer performance.

Test Order Sequence

A one-sample runs statistical analysis showed that the data were not influenced by subject test order sequence. The test order sequence changes in the data could have been caused by subject pretest learning generated by discussions between tested subjects and those waiting to be tested or by the experimenter unintentionally altering the test procedure or data collection techniques during the test. The one-sample runs analysis was applied to the data on the total number of errors for each subject. The analysis is based on the test order sequence in which the individual scores or observations were originally obtained. The analysis shows that the nine data runs are neither too many or too few for the six data above and 14 data below the average value, and therefore statistically nonsignificant (0.05 level) (Siegel, 1956).

Effects of Practice

The analysis shows that practice was not necessary on Day 4 and the subject would have obtained the same performance without it. However, the practice trials were necessary to improve performance during the speech-retention phase on Day 5 when enrollment training was not used.

Table 25 shows the effects of practice on performance for Day 4. The table shows that according to Friedman's test, there is no difference in the number of errors among the practice runs and the test runs. Table 26 shows the repeats needed for recognition given an initial mis-recognition for the practice and test runs of Day 4. As would be expected, a chi-square test shows no interaction between test runs and the number of repeats or differences by the repeats or test periods.

In contrast, Table 27 shows the effects of practice on performance for Day 5. The table shows significant difference in the number of errors for practice and test runs. A post hoc contrast test shows significantly more errors during the first five practice runs than during the last five practice and five test runs combined. Table 28 shows the repeats needed for recognition given an initial mis-recognition for the practice and test runs of Day 5. However, a chi-square test shows no significant differences.

Effect of Interruptions

The test session was interrupted when the subject ended his practice runs and began his test runs. There was a brief wait (several minutes) while the experimenter activated the VCR, checked camera alignment, and started the test program. However, the interruption of the task had no effect on performance. According to the binomial sign test, Table 29 shows that there is no significant difference between the last run of the practice session and the first run of the test session during Days 4 and 5.

Table 25

Friedman's Nonparametric Two-Way Analysis of Variance by Ranks for the Number of Errors for the Practice and Test Runs of Day 4

	Number of Errors			Ranking		
Subject	Pr4[a]	Pr4[b]	Test[c]	Pr4[a]	Pr4[b]	Test[c]
1	7	10	3	2	1	3
2	9	7	2	1	2	3
3	18	12	32	2	3	1
4	1	11	6	3	1	2
5	10	10	6	1.5	1.5	3
6	3	4	0	2	1	3
7	2	4	0	2	1	3
8	0	0	1	2.5	2.5	1
9	14	13	6	1	2	3
10	5	10	10	3	1.5	1.5
11	10	0	8	1	3	2
12	7	4	10	2	3	1
13	1	3	3	3	1.5	1.5
14	2	0	0	1	2.5	2.5
15	5	5	2	1.5	1.5	3
16	0	0	0	2	2	2
17	1	1	3	2.5	2.5	1
18	0	1	0	2.5	1	2.5
19	1	3	0	2	1	3
20	1	3	0	2	1	3
Total	97	101	92	39.5	35.5	45

Note. Friedman test value = 2.28, chi-square statistic (0.05: 2) = 5.99.

Day 4 (retention test) [a] - Runs 1-5, practice session
[b] - Runs 6-10, practice session
[c] - Runs 1-5, test session

Table 26

Number of Repeats Needed for Recognition Given Mis-Recognition for the Practice and Test Runs of Day 4

Test period	Number of repeats 1	2	3	4	5	6	>6
			Percentage				
Pr4[a]	71.64	17.91	5.97	2.99	1.49		
Pr4[b]	68.66	19.40	7.46	2.98	0.0	1.49	
Test[c]	82.80	10.30	5.20	0.0	0.0	0.0	1.70
			Frequency				
Pr4[a]	48	12	4	2	1		
Pr4[b]	46	13	5	2	0	1	
Test[c]	48	6	3	0	0	0	1

Chi-Square Test

Contingency Tables

Test period	Observed Combined repeats 1	2	3->6	Expected Combined repeats 1	2	3->6	Total
Pr4[a]	48	12	7	49.55	10.82	6.63	67
Pr4[b]	46	13	8	49.55	10.82	6.63	67
Test[c]	48	6	4	42.89	9.36	5.74	58
Total	142	31	19	142	31	19	192

Note. Test value = 3.52, chi-square statistic (0.05:4) = 9.49.

Day 4 (retention test)
[a] - Runs 1-5, practice session
[b] - Runs 6-10, practice session
[c] - Runs 1-5, test session

Table 27

Friedman's Nonparametric Two-Way Analysis of Variance by Ranks for the Number of Errors for the Practice and Test Runs for the Day 5 Retention Test Session

	Number of Errors			Ranking		
Subject	Pr5[a]	Pr5[b]	Test[c]	Pr5[a]	Pr5[b]	Test[c]
1	44	12	37	1	3	2
2	83	4	2	1	2	3
3	6	11	22	3	2	1
4	5	16	28	3	2	1
5	5	3	4	1	3	2
6	0	0	0	2	2	2
7	27	2	2	1	2.5	2.5
8	27	14	9	1	2	3
9	23	24	30	3	2	1
10	38	2	3	1	3	2
11	32	32	25	1.5	1.5	3
12	2	2	6	2.5	2.5	1
13	0	1	3	3	2	1
14	26	2	23	1	3	2
15	8	1	28	2	3	1
16	12	32	1	2	1	3
17	3	2	3	1.5	3	1.5
18	6	4	4	1	2.5	2.5
19	11	44	1	2	1	3
20	7	0	4	1	3	2
Total	365	208	235	34.5	46	40.5

Note. Friedman test statistic = 7.33, chi-square (0.05: 2) = 5.99.

Post Hoc Contrast Coefficients

Contrast	Pr5[a]	Pr5[b]	Test[c]	Value	Variance	Range
Pr5[a] vs Pr5[b], Test[c]	+2	-1	-1	-17.5	0.30	-18.8, -16.2
Pr5[a], Pr5[b] vs Test[c]	-1	-1	+2	0.5	0.30	-0.84, +1.84

Day 5 (retention test)

[a] - Runs 1-5, practice session
[b] - Runs 6-10, practice session
[c] - Runs 1-5, test session

Table 28

Number of Repeats Needed for Recognition Given Mis-Recognition for the Practice and Test Runs of the Day 5 Retention Test Session

Test period	Number of repeats 1	2	3	4	5	6	>6
			Percentage				
Pr5[a]	50.00	18.55	8.06	9.68	1.61	4.03	8.07
Pr5[b]	59.76	18.29	6.10	7.32	1.22	3.66	3.65
Test[c]	61.40	14.80	7.90	3.00	4.90	2.00	6.00
			Frequency				
Pr5[a]	62	23	10	12	2	5	10
Pr5[b]	49	15	5	6	1	3	3
Test[c]	62	15	8	3	5	2	6

Chi-Square Test

Test period	Combined number of repeats 1	2	3	4-5	6->6	Total
			Observed Contingency Table			
Pr5[a]	62	23	10	14	15	124
Pr5[b]	49	15	5	7	6	82
Test[c]	62	15	8	8	8	101
Total	173	53	23	29	29	307
			Expected Contingency Table			
Pr5[a]	69.88	21.41	9.29	11.71	11.71	124
Pr5[b]	46.21	14.16	6.14	7.74	7.74	82
Test[c]	56.92	17.44	7.57	9.54	9.54	101
Total	173	53	23	29	29	307

Note. Test value = 4.64, chi-square statistic (0.05:8) = 15.5.

Day 5 (retention test) [a] - Runs 1-5, practice session
[b] - Runs 6-10, practice session
[c] - Runs 1-5, test session

Table 29

Effects of Task Interruption on Performance Measured by the Number of Errors According to the Binomial Sign Test

	Test Session					
	Day 4 Practice test			Day 5 Practice test		
Subject	Run 10	Run 1	Sign	Run 10	Run 1	Sign
1	0	1	+	12	2	-
2	0	0	*	0	1	+
3	1	1	*	1	11	+
4	1	2	+	11	2	-
5	0	1	+	0	2	+
6	3	0	-	0	0	*
7	0	0	*	0	0	*
8	0	0	*	1	1	*
9	1	0	-	5	7	+
10	3	2	-	1	0	-
11	0	2	+	23	16	-
12	1	1	*	1	1	*
13	0	1	+	0	0	*
14	0	0	*	1	2	+
15	0	1	+	0	25	+
16	0	0	*	0	0	*
17	0	0	*	0	0	*
18	0	2	+	2	1	-
19	0	0	*	1	0	-
20	0	0	*	1	3	+
Note. Number signs pairs:			10			13
Number smaller group:			3			6
Binomial probability:			0.172			0.500

Task Variability Test

The occurrence of extreme mis-recognition numbers during the task variability test is similar to those of the Day 5 speech-retention test. The task variability test is a continuation of the speech-retention test with variable display rates, and the occurrence of multiple mis-recognitions may be because of no prior enrollment training. However, some of the multiple sequences for the first five runs of the task variability test are longer than those for the speech-retention test runs. These sequences decrease with test runs and eventually approach the level of the speech-retention test. Apparently, some of the subjects had to practice with the variable display update rate before they could return to the performance level of the previous test.

The mis-recognition data for the task variability test (see the Appendix) show that the increased variance in the mis-recognitions are because of extreme runs of multiple mis-recognition sequences scattered among the subjects and test runs. The table shows that while the incidents of extreme mis-recognition sequences are scattered uniformly through the test runs and were experienced at least once by 40% of the subjects, the number in the sequences decreases with increasing test run. The table also shows that the occurrence of extreme mis-recognition values is similar to those of the Day 5 speech-retention test.

Table 30 lists the average number of errors and standard deviation for the test runs in groups of three; most of the subjects experienced all three display rates in random order across three consecutive runs. The table shows a decrease in the average number of errors as the number of groups increase. The standard deviation is not as well behaved; however, the application of the Hartley test shows that the variance of the first group is significantly greater than the pooled variance of the remaining groups. The average and standard deviation are plotted in Figure 11 for the test runs in groups of three. This figure demonstrates the large deviation in the errors for the initial three test runs, followed by the narrowing of deviations about the averages as the subjects adjusted to the variable update rate.

Table 30 also lists the total number of errors for the three rates as a function of the test order during which the rates were seen by the subjects. The data suggest a decrease in the number of errors with test order at least for the two higher rates; however, the trend is nonsignificant as shown by the application of a nonparametric Friedman ANOVA test by ranks.

Apparently, the effect of introducing the change in display update rate was to initially sharpen the performance of some subjects while degrading the performance of others. The degradation was produced by the occurrence of mis-recognition runs. As the subjects practiced with the changed conditions, their performance returned to the previous levels. In this way, the variation of response was initially increased while the average response remained unchanged.

Subject Errors

The subjects committed very few errors. This is surprising considering the large number of mis-recognitions. Apparently, the task was simple enough that the large number of mis-recognitions did not confuse the subjects about the correct task to be performed. The total errors made by the subjects as they performed the tasks are listed in Table 31 (as a function of test practice sessions and of test runs). A subject error is defined as a departure from the expected sequence of actions needed to perform the task in an optimal manner. This includes selecting the wrong instruction menu, selecting menus out of sequence, or using the wrong voice command.

Military Training and Experience

Of interest is a possible explanation of the recognizer performance in terms of the characteristics of the subjects as determined from the demographic data listed in Table 5. Certainly, the subject's experience and training in giving speech commands may influence the automatic speech recognizer performance. These factors may be measured by the military occupational specialty (MOS) and military rank. Table 5 shows that all subjects had reached the same educational level. All were high school

Table 30

Statistics for the Number of Errors for the Day 5 Task Variability Test

A. Average and standard deviation for the test runs in groups of three

	Test Runs				
Parameter	1-3	4-6	7-9	10-12	13-15
Average	3.60	2.27	2.87	2.45	2.17
Standard deviation	9.83	3.08	6.38	5.07	4.49

B. Hartley's test of the homogeneity of variance for test runs 1-3 and the remaining test runs polled

Test value = 4.02
Hartley's test statistic, Fmax(.05: 2,20) = 2.46

C. Number of errors for display update rate as function of test order

	Test Order				
Rate	1	2	3	4	5
X1	34	61	71	58	26
X2	65	52	68	40	42
X3	117	23	33	49	62

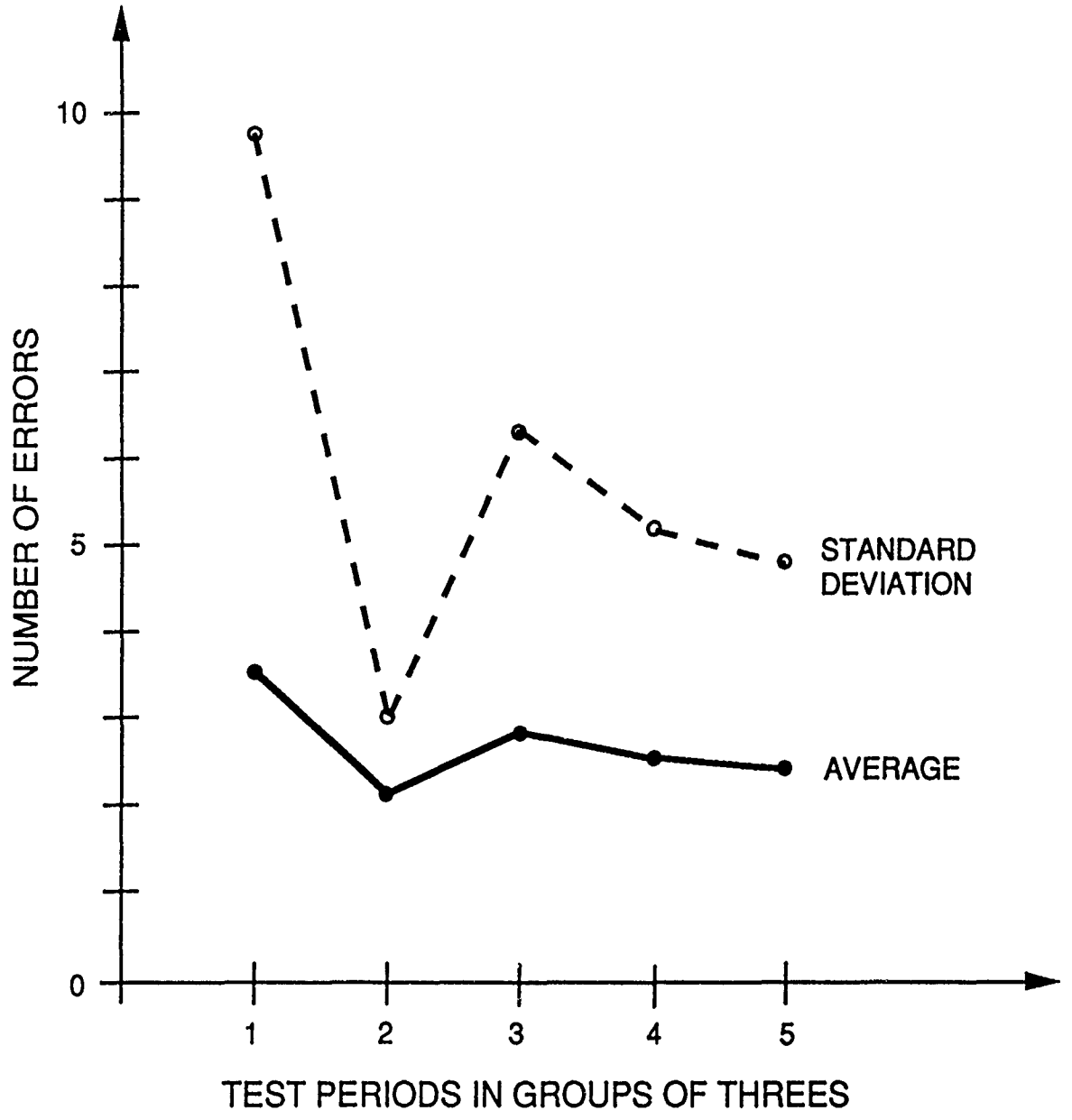

Figure 11. Number of errors statistics for the task variability test by test runs in groups of three.

Table 31

Subject Errors by Test Run for Test and Practice Sessions

Session	Run 1	Run 2	Run 3	Run 4	Run 5	Total
Day 1	3	1	3	0	2	9
Day 2	1	1	0	1	0	3
Day 3	0	0	0	0	0	0
Pr4[a]	2	3	0	0	0	5
Pr4[b]	0	1	0	1	1	3
Day 4	1	0	2	1	1	5
Pr5[a]	2	0	0	0	0	2
Pr5[b]	0	0	0	0	0	0
Day 5	0	0	0	5	0	5
Rate 1	0	0	1	4	2	7
Rate 2	0	2	2	5	0	9
Rate 4	0	1	3	0	3	7
Total	9	9	11	17	9	55

a - Runs 1-5, practice session
b - Runs 6-10, practice session

graduates; a few had 1 year of college. A study of the data shows a high correlation between military rank and subject age, with the older subjects being of higher rank. In the following paragraphs, the relationship between automatic speech recognizer performance, military rank, and MOS is analyzed.

Interaction of Military Rank and Military Occupational Specialty

A study of subject distribution by military rank and MOS shows that the infantry subjects (MOS 11B/11M) and tank armor crewmen (MOS 19E/19K) have higher ranks, while most of the cannon crewmen (MOS 13B) and wheeled vehicle operators (MOS 88M) have lower ranks (see Table 32, Part A). The total number of errors data shows that while the lower ranking wheeled vehicle operators and tank armor crewmen made more errors than their higher ranking counterparts, the converse is true for the infantrymen and artillerymen. However, the number of subjects per rank by MOS cell is too small for random samples and will not support a proper statistical analysis of the interaction. This remains true even after reducing the interaction matrix by combining MOSs and ranks. For this reason, separate analyses by ranks and MOS follow.

Table 32

Chi-Square Analysis of the Number of Automatic Speech Recognizer Errors by Military Rank and Military Occupational Speciality

A. Number of automatic speech recognizer errors for each subject listed by military rank and military occupation specialty

	MOS			
Rank	88M	19K/E	11B/M	13B
E2	19			
	13			
	102			
E3	304	62		8
				6
E4	48	118	15	105
	5		38[a]	
	27			
E5	30	91	26[a]	
			5[a]	
			19[a]	
E6		12		

[a] - Subject from 1987 test.

B. Chi-square analysis of the average number of automatic speech recognizer errors in term of the military occupation speciality grouped by wheeled vehicle operators and the combat arms

	MOS	
	Operators	Combined arms
Observed	68.5	42.1
Expected	55.3	55.3

Note. Chi-square test value = 6.31, statistic (.05,1) = 3.84.

Table 32 (Continued)

C. Chi-square analysis of the average number of automatic speech recognizer errors by military rank

	Rank		
	E2-E3	E4	E5-E6
Observed	73.4	50.8	27.5
Expected	50.6	50.6	50.6

Note. Chi-square test value = 20.84, statistic (.05,2) = 5.99.

Military Occupational Specialty

A nonparametric chi-square analysis of the total number of errors by subject shows that the combat arms group, consisting of the infantrymen (MOS 11B/11M), cannon crewmen (MOS 13B), and tank armor crewmen (MOS 19E/19K), had statistically significant fewer errors than did the wheeled vehicle operators (MOS 88M) (see Table 32, Part B). These results are expected since all combat arms specialties have extensive experience in communications, either by internal telephone linkage with crewmen within their vehicles or units or by radio linkage with outside sources in the command and control network. The chi-square analysis is based on the average number of errors per subject in the two specialist groups (since the number of subjects is different with eight subjects in the wheeled vehicle group and 12 subjects in the combat arms group). However, the results are questionable since the groups are not balanced by rank; most of the wheeled vehicle group are lower rank, while most of the combat arms group are higher rank.

Military Rank

A nonparametric chi-square analysis of the total number of errors by subject shows that the subjects with higher military rank had statistically significant fewer errors than did those with lower rank (see Table 32, Part C). The number of errors is largest for E2 to E3 ranks, smaller for the E4 rank, and smallest for E5 to E6 ranks. The chi-square analysis is based on the average number of errors per subject in each rank grouping. The number of subjects varies for the different groupings with six subjects in the E5 to E6 ranks and seven subjects each in the E2 to E3 and E4 ranks. The results are as expected since the higher ranks tend to be older, have more experience in military systems, are trained in giving commands at the supervisory task levels, and have confidence that their commands will be followed. However, as noted previously, the rank categories are not balanced by MOS since most of the lower ranks are in a vehicular MOS.

Effects of Prior Experience

Four subjects in this test participated in the 1987 test (Smyth et al.) and therefore had some prior experience with the automatic speech recognizer and test procedure. These subjects had no difficulty with the automatic speech recognizer in the 1987 test and performed at a high recognition rate in that test. However, they had no other experience with speech recognition systems.

As before, these subjects had little difficulty with the automatic speech recognizer and in this study performed at a high recognition rate. These subjects were higher ranking infantrymen and their high performance may be the reason for the group's high performance (see Table 32, Part A). Furthermore, since they volunteered for this test (no one who had performed poorly in the 1987 test did so), their results may have biased the test results to appear more favorable than they would have with the users population in general.

Interestingly, their dress, bearing, attitude, and speech delivery were distinctively outstanding. Their speech was robust in sound quality and their words were distinct. They were confident that they would be understood and their orders would be followed. There was also an undercurrent of competition in their behavior. In short, they were team-playing "winners" who maximized their opportunities to excel. However, a review of the data shows that their error count was about the same as that of other subjects who had performed well during testing, regardless of rank and MOS.

Nevertheless, these observations suggest that high recognizer performance may be correlated to an innate mannerism of speech which is a reflection of personality, natural intelligence, and physical characteristics. Individuals having these characteristics may be attracted to a more challenging combat arms position in a communication-driven organization, such as the peacetime Army. These soldiers may be the ones promoted. unfortunately, the data are too sparse to confirm this observation.

Subjects' Ratings

On a scale of 1 through 10, the subjects were asked to rate the automatic speech recognizer's usefulness to the Army. Most subjects rated the system relatively high. Furthermore, most subjects' ratings improved after training test sessions (Days 1 through 4). Subjects and test sessions (Days 1 through 5) are listed in Table 33. This table shows that the average rating tracks an increase in performance as training improved performance and a decrease during the Day 5 retention test.

Observations

As mentioned previously, four subjects participated in the 1987 study (Smyth et al.) and therefore had experience with the command vocabulary and with using the automatic speech recognizer in performing the task. These subjects had no difficulty with the automatic speech recognizer during the 1987 test and maintained high recognition accuracies. Because of their previous success, they readily volunteered for this test. One subject was a member of Group I; the remaining three were members of Group II. One subject had difficulty during the speech-retention test on Day 5; otherwise, all four subjects performed well. The conclusion is that subjects who do well with the automatic speech recognizer can continue to perform well years later.

Table 33

Subject Ratings From One to Ten About the Usefulness of the Automatic Speech Recognizer to the Army

	Test Session				
Subject	Day 1	Day 2	Day 3	Day 4	Day 5
1	x	8	8	8	8
2	7	8	9	9	9
3	10	7	6	6	6
4	6	6	6	6	7
5	5	7	6	x	7
6	7	7	7	7	7
7	8	8	8	8	8
8	9	9	8	8	6
9	8	8	8	7	7
10	9	9	9	9	9
11	5	6	10	10	10
12	9	9	9	9	9
13	10	10	10	10	10
14	9	9	9	9	9
15	8	8	8	8	7
16	8	7	10	10	x
17	8	8	9	9	9
18	8	8	8	8	8
19	8	9	9	9	9
20	5	9	10	10	10
Average	7.74	8.00	8.35	8.42	8.16
SD	1.56	1.08	1.31	1.2[illegible]	1.30

Note: x - datum not recorded.

It is interesting to note, however, that no one who had difficulty with the automatic speech recognizer during the 1987 test volunteered to participate in this study. Therefore, another conclusion is that the user population, as with any application, will tend over time to be represented by those who can successfully use this device. Unfortunately, these are not necessarily all the members of the intended user population.

Some subjects who initially performed poorly improved by learning to maintain consistent voice patterns for both enrollment and task. These subjects apparently became conscious of their voice inflections and developed strategies for voice discipline. In fact, several subjects mentioned that they had been challenged by the test to develop better voice control.

Motivation was a decisive factor in the subject's performance. Most subjects who initially performed poorly felt challenged to improve during the following sessions. One subject reported that he enjoyed the study because he learned how to control his speech delivery to the point that the automatic speech recognizer became a verbal typewriter which he operated with the finesse of his voice. Subjects who initially did very well (more than 90% recognition) rapidly lost interest in the study; however, their performance decreased only slightly. Some subjects, who initially did well, had no need to develop voice discipline strategies and were discouraged when first encountering mis-recognition sequences in the speech-retention phase. Finally, some subjects who initially performed only moderately well (80% through 90%) felt no need to improve and did not strive to develop voice discipline strategies.

CONCLUSION

The results of this study show that training produced only a slight improvement in automatic speech recognizer performance for most of the subjects. These subjects had little or slight difficulty with the automatic speech recognizer and performed at a high machine recognition rate. However, a large minority (35%) of the subjects had a difficult time training their voices to be machine recognizable. These subjects required several days of training before their recognition rates were comparable to the majority of the subjects; even then, they never reached the performance level of the other subjects.

The subjects who were helped by training may have needed training in several different areas before being able to perform the task at their best ability. Unfortunately, the experimental results do not show what the subjects needed to learn. Certainly, some subjects needed to learn the enrollment procedure for the automatic speech recognizer. The procedure was long and tedious. Although the enrollment procedure was carefully controlled, some subjects' utterances during the first training session were not synchronized with the training prompts, thereby possibly mistraining the automatic speech recognizer. Furthermore, the subjects had to learn the task, voice commands, and use of the automatic speech recognizer in controlling the display. Finally, some subjects had to learn speech discipline in which the same speech pattern was used during enrollment, task training, and testing.

Considering the many factors influencing the performance of automatic speech recognizers (as noted in this report and others), the author recommends that reports about automatic speech recognizer experiments and results be standardized to include the following: the number of subjects, selection of subjects, type and length of training, enrollment procedures, dismissal of subjects, mis-recognitions, and number of mis-recognition runs.

In light of this study's results, the author recommends caution in the implementation of present-day automatic speech recognition systems in Army combat systems. The author supports future research into the development and performance of these systems in operational environments. A possible area for implementation of present technology is in the noncombat areas of logistics, quality control, and inventory systems.

RECOMMENDATIONS FOR FURTHER RESEARCH

The major source of recognition errors is the multiple sequences of nonrecognitions. The conclusion is that subjects learned to control their speech patterns to ensure consistency between the enrollment and task procedures. The errors are the result of differences between the speech patterns. It is recommended that the recorded speech patterns be analyzed for pitch, intensity, and timing of delivery to confirm a difference between the patterns recognized and those that were nonrecognized.

One result of this study is the importance of task-related training during an adaptation period before usage. An improvement in training could be incorporating the enrollment procedure into a task-related structure, thereby giving meaning to the voice commands. Furthermore, a visual feedback during the enrollment process showing a frequency spectrum or intensity time line could allow the user to compare his utterances to either a standard or his past history.

It is recommended that future developments in speech recognition technology be tested as they are developed for application in Army systems. This research is necessitated by the limited technology of today's automatic speech recognizer systems. It should come as no surprise that the performance of a speaker-dependent system is influenced by speech command deviations in pitch, intensity, and duration from those of the machine training set. Furthermore, it should be expected that the persistence of a speaker in maintaining consistent speech is determined by his experience in giving verbal commands and the task to be performed. One may expect the future development of robust speech recognition technology to be improved to the point of both speaker and task independence. However, it is only through testing (as reported here) that one can confirm this development.

REFERENCES

Billingsley, P. A. (1982). Navigation through hierarchical menu structures: Does it help to have a map? Proceedings of the Human Factors Society 26th Annual Meeting, 103-107.

Department of Defense. (1981). Military standard on human engineering design criteria for military systems, equipment, and facilities (MIL-STD-1472C). Washington, DC: U.S. Government Printing Office.

Department of Defense. (1983). Military standard on symbols for army air system displays (DoD-STD-1477)(MI). Washington, DC: U.S. Government Printing Office.

Dotson, D. A., & Smyth, C. C. (1987). Comparison of techniques for hooking tracks on a forward area air defense command and control (FAAD C2) display I (Technical Memorandum 9-87). Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

Fallesen, J. J., Smyth, C. C., & Blackmer, R. F. (1983). Human engineering laboratory division air defense systems I (HELDADS-I): Baseline air battle management operations center (ABMOC) manual performance (Technical Memorandum 11-83). Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

French, B. A. (1983). Some effects of stress on users of a voice recognition system: A preliminary inquiry. Monterey, CA: Naval Postgraduate School.

Howie, J. H. (1987). Speech technology in automotive assembly applications. International Speech Technology 87, Voice Input/Output Applications Conference. New York: Media Dimensions.

Malkin, F, J., & Christ, K. A. (1985). A comparison of voice and keyboard data entry for a helicopter navigation system (Technical Memorandum 17-85). Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

Marascuilo, L. A., & McSweeney, M. (1977). Nonparametric and distribution-free methods for the social sciences. Belmont, CA: Brooks & Cole.

Martin, J. D., & Poock, G. K. (1984). An initial look at stress and voice recognition. Journal of the American Voice Input/Output Society, 1(1), 24-33.

Miller, D. P. (1981). The depth/breadth tradeoff in hierarchical computer menues. Proceedings of the Human Factors Society 25th Annual Meeting, 542-546.

Pallett, D. S. (1985). Performance assessment of automatic speech recognizers. Journal of Research of the National Bureau of Standards, 90(5), 371-387.

Poock, G. K., & Martin, B. J. (1984). Effects of emotional and perceptual-motor stress on a voice recognition systems accuracy: An applied investigation (NPS 55-84-002). Monterey, CA: Naval Postgraduate School.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. New York: McGraw Hill.

Simpson, C. A., McCauley, M. E., Roland, E. F., Ruth, J. C., & Williges, B. H. (1985). System design for speech recognition and generation. Human Factors, 27(2), 115-141.

Simonov, P. V., & Frolov M. V. (1973). Utilization of human voice for the estimation of man's emotional stress and state of attention. Aerospace Medicine, 256-258.

Smyth, C. C., Denny, S. M., & Dotson, D. A. (1987). Comparison of voice recognition, touch panel, and keypad rechniques of data entry for a forward area air defense command, control, and intelligence (FAAD C2I) display (Technical Memorandum 25-87). Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

Velleman, P. F., & Hoaglin, D. C. (1981). Applications, basics, and computing of exploratory data analysis. Boston: Duxbury Press.

Verbex. (1983). Verbex series 4000 voice planner software user's guide. Bedford, MA: Verbex, Exxon Enterprises, Exxon.

Winer, B. J. (1971). Statistical principles in experimental design. New York: McGraw Hill.

APPENDIX

MIS-RECOGNITION DATA

Table 1

Mis-Recognition Errors for Test Sessions as a Function of Subjects and Test Runs

	Day 1 Test Runs					Day 2 Test Runs					Day 3 Test Runs				
Subject	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	1	5	12	0	0	0	1	1	1	1	2	4	2	3	2
2	50	0	50	51	53	7	9	14	10	57	0	0	0	1	0
3	5	1	0	3	0	0	1	1	0	0	3	0	0	0	2
4	3	4	47	11	1	1	0	0	0	0	7	5	2	1	3
5	3	3	9	13	20	1	1	2	1	0	1	0	1	0	1
6	8	17	58	26	8	0	0	0	0	0	0	1	0	0	0
7	0	0	1	1	2	1	0	0	0	1	0	1	0	0	0
8	1	0	0	0	1	0	0	2	0	0	0	0	4	1	1
9	0	0	1	2	1	1	2	1	0	0	0	0	0	0	0
10	0	0	0	0	0	0	1	0	1	0	2	0	2	1	0
11	27	22	5	11	12	1	1	0	0	0	0	1	0	0	0
12	1	0	0	0	0	8	1	0	3	0	3	1	3	3	5
13	0	0	0	2	0	1	0	1	2	0	0	1	2	0	0
14	1	0	0	0	0	0	0	1	0	0	1	1	3	0	1
15	1	0	0	1	1	0	0	0	5	2	0	3	0	0	0
16	0	0	0	0	0	3	0	1	2	2	0	0	0	0	0
17	0	0	6	0	3	1	47	21	19	2	0	2	3	1	1
18	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0
19	1	2	7	4	19	1	1	1	0	1	1	0	0	0	0
20	20	0	2	1	0	0	6	1	0	0	0	0	0	0	0

Table 2

Mis-Recognition Errors for Test Sessions as a Function of Subjects and Test Runs

	Practice					Practice					Test				
Subject	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5
							Day 4								
1	0	3	0	0	4	7	2	0	1	0	1	0	4	11	0
2	1	8	0	0	0	3	0	4	0	0	0	1	1	0	0
3	0	4	6	4	4	2	0	3	6	1	1	27	1	1	6
4	0	0	0	1	0	0	4	0	6	1	2	0	0	0	4
5	1	1	1	2	5	1	1	1	7	0	1	1	1	0	3
6	0	3	0	0	0	0	0	1	0	3	0	0	0	0	0
7	0	2	0	0	0	0	0	4	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	4	2	1	6	1	2	1	5	4	1	0	2	3	1	0
10	1	1	0	2	1	3	1	1	2	3	2	3	1	4	0
11	1	2	2	4	1	0	0	0	0	0	2	1	2	3	0
12	1	0	2	3	1	1	0	1	1	1	1	2	2	5	0
13	0	1	0	0	0	0	0	2	1	0	1	0	1	0	1
14	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
15	0	1	1	1	2	2	1	1	1	0	1	0	0	1	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	1	0	0	0	2	0	1	0	0	0	0	0	0	0
18	1	0	0	0	0	0	0	0	1	0	2	0	1	0	0
19	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
20	1	0	0	0	1	0	0	0	0	0	0	0	0	1	2
							Day 5								
1	0	5	0	2	5	4	1	9	8	12	2	10	8	4	13
2	17	20	44	0	1	0	4	0	0	0	1	0	0	1	0
3	1	3	1	1	0	0	0	10	0	1	11	8	0	1	2
4	1	3	1	0	0	0	1	0	4	11	3	0	20	2	3
5	2	1	1	1	0	0	1	2	0	0	2	0	1	1	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	5	2	4	13	3	0	0	0	2	0	0	0	1	0	1
8	3	0	0	0	4	6	7	0	0	1	1	0	3	0	5
9	3	11	3	5	1	3	7	4	5	5	7	1	6	5	11
10	36	1	0	1	0	0	0	0	1	1	0	0	1	2	0
11	23	4	4	1	0	0	3	3	3	23	16	3	2	4	0
12	0	0	1	0	1	0	1	0	0	1	1	2	0	1	2
13	0	0	0	0	0	0	0	0	1	0	0	0	0	3	0
14	9	6	7	1	3	0	0	0	1	1	2	2	0	4	15
15	0	0	0	0	8	1	0	0	0	0	25	1	0	0	2
16	1	0	2	8	1	2	2	1	27	0	0	0	1	0	0
17	0	0	1	5	1	0	0	0	0	0	0	0	0	1	0
18	3	0	0	0	0	0	0	0	0	2	1	1	0	0	1
19	1	3	0	1	1	1	2	0	0	1	0	1	1	0	2
20	6	0	2	1	2	30	4	1	8	1	3	1	0	0	0

Table 3

Mis-Recognition Errors for Test Sessions as a Function of Subjects and Test Runs (r)

	Task Variability Test Runs														
Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	3	7	7	9	7	6	30	16	1	4	5	0	3	3	12
2	0	1	0	0	0	0	0	0	0	0	2	0	0	1	0
3	0	2	2	0	2	2	2	4	5	1	1	1	10	1	0
4	0	4	1	3	0	0	0	0	0	0	0	1	0	0	0
5	1	0	1	2	1	0	1	2	0	1	2	1	2	1	2
6	1	1	0	3	0	0	0	0	0	1	0	1	1	1	0
7	0	1	0	3	0	0	9	1	0	0	0	0	0	2	1
8	2	62	6	10	6	0	2	0	1	0	0	1	0	0	0
9	17	25	3	13	6	6	6	3	12	6	7	7	6	3	25
10	2	3	2	3	5	2	1	2	2	4	4	6	9	6	4
11	3	7	38	2	7	5	14	34	3	35	11	4	9	19	11
12	0	0	0	1	1	1	0	2	0	0	0	1	0	0	2
13	0	1	1	1	0	0	0	1	0	0	2	0	0	0	0
14	1	0	2	0	2	0	0	0	0	1	1	1	1	0	0
15	0	1	0	1	1	0	0	0	1	0	0	0	3	0	1
16	0	2	0	0	11	0	1	0	1	4	0	0	1	0	0
17	2	0	0	1	1	0	0	1	1	0	1	0	1	2	2
18	1	1	0	4	0	0	1	1	0	0	1	0	1	0	2
19	1	0	1	1	2	1	1	3	4	1	3	1	2	3	0
20	0	0	1	4	0	0	0	1	2	0	0	1	0	0	0