The Effects of 16 Variables on a Telephone Information System which Uses Synthetic Speech

by

Douglas Barrett Beaudet

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Industrial Engineering and Operations Research

APPROVED:

Robert C. Williges, Chairman

Beverly H. Williges

Dennis L. Price

October, 1988

Blacksburg, Virginia

The Effects of 16 Variables on a Telephone Information System which Uses Synthetic Speech

by

Douglas Barrett Beaudet

Robert C. Williges, Chairman

Industrial Engineering and Operations Research

(ABSTRACT)

Information systems that employ synthetic speech are emerging daily in the consumer market. However, many of these systems are being developed without first investigating the numerous factors that affect the design and usability of these systems. This study investigated the effects of 16 variables on a telephone information system which uses synthetic speech as the display modality. The information system was for a fictitious department store. Subjects telephoned the system and searched for information messages on specific store items. Upon hearing the message, subjects transcribed what they heard and rated their perceived difficulty in understanding the message, their confidence in correctly remembering the message, and their perceived difficulty in finding the store item in the system. Subject search performance measures were recorded during each search, and system evaluation subjective ratings were collected at the end of each experimental session.

A Hadamard 32x32 matrix design was used in this screening study to test efficiently the main effects of the 16 variables on 23 measures of user performance. Only 32 data points were required to evaluate the variables in the screening study. The analyses identified 8 variables (speech rate, menu organization, number of

targets, wallet guide, menu feedback, background music, subject age, and subject gender) as having a significant effect in at least two tests; 4 variables (voice type, pause/resume, repeat keyword, and command feedback) as having a significant effect in one test; and 4 variables (input timeout, system response time, selection feedback, and spell-out keyword) that did not have a significant effect in any test. The analyses also assessed the worth of the 12 dependent measures in providing meaningful test results.

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation on contract number IRI 8604793. Dr. H. E. Bamford Jr. served as technical monitor. The co-principal investigators on the contract were Dr. Robert C. Williges and Beverly H. Williges. I am truly indebted to the Williges' for the opportunity to participate in this research, as well as the education I received from them. I also wish to thank Dr. Dennis Price for his insightful contributions and recommendations which greatly enhanced this study. Thanks goes to Dr. Robert Dryden for his assistance in recruiting "willing and able" subjects. Calvin Selig, System Manager for the Virginia Tech Human Computer Interaction Laboratory, developed the software for this project, and without him I would still be sitting in the Lab. I would also like to thank my office-mates and co-workers, Peter Jay Merkle Jr. and David W. Herlong; two of the best guys in the world to have around when you are looking for answers.

I want to thank my parents, David Walter and Shirley Taylor Beaudet, for a lifetime of education that brought me here. Most of all, I want to thank my lovely wife Deena, who gave me inspiration to go back to school, motivation to stay in school, and conviction to see it through to the end.

DEDICATION

To Julie Michele Beaudet, my little Miss Magic.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

INTRODUCTION

This screening study investigated the effects of 16 variables on a telephone information system which uses synthetic speech as the display medium. The 16 variables were identified during an engineering analysis of telephone information systems (Merkle, 1988) as having a high probability of affecting user performance. The purpose of this study was to provide preliminary findings on the main effects of the 16 variables and direction for future research in the area of telephone information systems which use synthetic speech.

In this study, an information system for a fictitious department store was studied. Subjects telephoned the system and searched for information messages on specific store items. Upon hearing the message, subjects transcribed what they heard and rated the perceived difficulty of understanding the message, their confidence in correctly remembering the message, and the difficulty of locating the store item in the database. Subject search performance measures were scored against calculated expert performance scores. Transcription accuracy and subjective ratings were recorded and analyzed.

A Hadamard 32x32 matrix design (Diamond, 1981) was used to investigate the 16 variables economically. The design used a minimum number of data collection points to evaluate the main effects of all 16 variables. For the purposes of this study, only main effects were investigated to provide preliminary data on what variables have a significant effect on user performance.

Variable interactions were not considered in this study because the objective of the screening study was to investigate efficiently numerous factors with a small number of data points. Screening studies are intended to be used to identify

important factors, and not to obtain an accurate representation of the experimental space for the variables under consideration (Simon, 1977). Variable interactions can be investigated in follow-on studies for those variables identified in the screening study as having a significant main effect on user performance. Therefore, for this screening study, the effects of variable interactions are considered to be non-significant.

Information Systems

Information systems provide specific information on selected items for a defined subject area. Telephone books, operator directory assistance, library card catalogs, computerized library systems, department store catalogs, and automated banking machines are all examples of information systems. With the advent of the computer age, such systems have become very popular and worthwhile. The data storage capabilities of computers have permitted these Socratic systems to develop almost overnight.

Today, many consumer companies have telephone information systems to handle customer inquiries. These systems are typically operated by service operators stationed at computer terminals which are connected to product information systems. Companies have found information systems to be cost effective for the company and convenient for the customer. In theory, large amounts of information are easily accessible and readily available; thus, satisfying company and consumer needs. Naturally, the assumption is that these systems have been designed for effective user interaction.

In the race to build a better information system, attention is now focused on incorporating synthesized speech as the delivery medium for information. The goal is to replace the mundane job of the human operator (who searches, retrieves, and speaks repetitious or picayune information) with a speech synthesizer. Recent improvements in speech technology have made synthetic speech a viable alternative to the human operator for many types of information delivery systems. However, synthetic speech is not a cut-and-paste solution. The technology has limitations, as well as numerous factors which should be considered before applying it to any information system.

In the past ten years, synthesized speech has been applied to a number of information systems including speech filing systems for voice messages (Gould and Boies, 1983; Schmandt, 1985a); an information service system (Podgorny, 1985); a tourist information system for the city of Austin, Texas (Thomas, Rosson, and Chodorow, 1984); and a telephone inquiry service for accessing inventory stores, games, and message systems (Witten and Madams, 1977). The majority of these applications were developed through designer intuition or personal preferences. Such design tactics often result in a final product which is, at best, a less than optimum design. Often, inconspicuous design flaws manifest themselves during usability testing or well into production and marketing.

This screening study is part of an overall research project investigating telephone inquiry systems which use synthesized speech. The research project methodology is based on the theory of sequential research strategies (Cochran and Cox, 1957). While the basic application of using synthetic speech in an

information system has met with limited success, there remains a large amount of research which needs to be performed in this area.

Research Paradigm

The research project, under which this research was conducted, is the Integrated Research Paradigm for Experimentation in Information Technology and is supported by a research grant from the National Science Foundation. The objective of the project is to develop a methodology for investigating design issues associated with new and complex information systems. Researching telephone information systems serves as an experimental paradigm for demonstrating the methodology.

Information systems have a multitude of human, task, hardware, software, and environmental issues which must be considered during system development. Historically, researchers and systems designers have conducted research on very specific interactions between certain design issues. This has resulted in small pockets of information, which were usually unrelated to one another. As a result, there has been an inability to generalize research data to aid in the design of systems.

The research project methodology for this project uses sequential research strategies (Cochran and Cox, 1957) for investigating the experimental paradigm. There are three major components to the sequential research strategy; (1) select the most important variables from the total variable set for investigation, (2) describe the functional relationships of these variables in terms of an empirical model, and (3) assess the optimum configuration for the experimental paradigm.

In Step One, early investigations identified 95 independent variables which could affect user performance for a telephone information system which uses synthesized speech. An engineering analysis was performed to identify the critical variables for investigation in Step Two. The engineering analysis consisted of reviewing the relevant literature, obtaining subjective evaluations from users and human factors experts of a prototype system, and performing a feasibility study for manipulating each variable.

A series of working group meetings was held by the research team (five human factors engineers) to analyze the data collected. Rules were developed and employed for each data source to identify the most salient variables. In the end, 16 out of 95 variables were selected to be investigated in a screening study.

This study represents the first experimentation in Step Two. The results of this screening study provide direction for the other experiments to be conducted in Step Two. Step Three will be performed by using response surface methodology to link together the results of Step Two and describe the experimental space for the research paradigm.

Experimental Overview

The variables identified for the screening study address design issues involving voice structure, database organization, system control rates, command features and feedback, user aids, and environmental considerations. In Table 1, each of the 16 variables is listed by design issue.

Table 1 Screening Study Independent Variables

Variable	Levels
Speech Display	
type of voice	P. Paul, B. Betty
speech rate	180, 240 wpm
Database	
menu organization (bxd)	8x2, 2x6
number of target items	1, 2
System Control Rates	
input timeout	2, 4 sec.
system response time	0, 4 sec.
Command Features	
pause/resume	available, not available
repeat keyword	available, not available
spell-out keyword	available, not available
User Aids	
wallet guide	available, not available
menu selection feedback	available, not available
database selection feedback	available, not available
command selection feedback	available, not available
Environment	
background music	none, background
User Demographics	
age of subject	18-30, 45-60
sex of subject	male, female

Purpose

The purpose of this study was to test the effects of 16 independent variables on 23 dependent measures. The screening study determined whether any of the 16 variables has a main effect on user performance with a telephone information system which uses synthetic speech. In addition, the screening study also determined the usefulness of each of the 23 dependent measures. By evaluating the usefulness of these dependent measures, conclusions have been reached as to the appropriateness of assessing these measures in future studies.

LITERATURE REVIEW

Speech Synthesis

Speech synthesis is an auditory display which has the fundamental characteristic of sounding similar to human speech. As with other auditory displays, synthesized-speech can be effective only if it is heard and understood by its users. Synthesized speech has unique characteristics which make it well suited for certain applications and inappropriate for other applications. Guidelines exist for using synthesized speech for informative or emergency displays (Sanders and McCormick, 1987). However, little research has been conducted on applying speech synthesis in interactive communication or information systems (Simpson et al., 1985).

Speech synthesis technology. Speech synthesis systems convert strings of computer output code into audio signals which communicate information by imitating human speech. Typically, speech synthesis systems are developed for specific applications in a defined subject area and are not designed to communicate information outside of this defined subject area.

There are two methods for generating synthetic speech; synthesis-by-rule and digitized human speech. Synthesis-by-rule uses algorithms containing rules of pronunciation for a given language to produce speech directly from strings of computer code. Rules of phonemes (sounds of speech) provide the necessary characteristics of speech duration, energy, accent and resonance to imitate a human voice. Additional rules provide direction for eliminating ambiguity between similar words. A complex combination of filters and resonators model the vocal tract to

produce quasi-human speech sounds. Synthesized speech can be either male or female, young or old; manufacturers typically offer from one to six voice types.

Digitized human speech uses a recording of a person's voice to generate speech. This is accomplished by either digitally recording a human speaker saying the desired words (digitized speech) or speaking a series of human sounds (synthesis-by-analysis). For digitally recorded speech, the recording usually is a tape recording of a human speaker saying all the possible words which may be spoken by the machine. This technique provides high quality speech generation and is technically simple to record, but is very costly to manipulate (Schmandt, 1985b).

For synthesis-by-analysis, a computer records a series of human sounds as sample speech waveforms. The recordings are compressed in format using techniques such as Fourier Transform, Linear Predictive Coding, or Waveform Parameter Encoding (Schmandt, 1985b; Simpson, et al., 1985). To generate speech, computer output text is matched with corresponding waveforms to generate words and phrases. The quality of digitized speech is largely dependent upon the fidelity of the digital recording -- digitization rate and original signal quality.

Evaluating speech synthesizers. Speech synthesizers are judged for user acceptability along two parameters: intelligibility and naturalness. Intelligibility can be interpreted as the percentage of single word utterances recognized correctly or words in a sentence which are recalled correctly. Intelligibility is influenced not only by the articulation of the synthesizer per se, but also by word length, sentence structure and length, and speech rate (Merva, 1987; Simpson and Marchionda-Frost, 1984). In addition, Merva (1987) found that transcription accuracy was

significantly better for the words at the end of a message than for the words at the beginning of a message.

Naturalness is the listener's subjective rating of how much the generated speech sounds like human speech. Naturalness of synthesized speech has a significant effect on user acceptability (Rosson and Cecala, 1985; Simpson and Marchionda-Frost, 1984). However, for some applications users prefer the synthesizer to have an artificial or "mechanical" sound (Simpson and Marchionda-Frost, 1984). Simpson and Marchionda-Frost reported that in the confines of the helicopter cockpit, pilots preferred that the synthesizer voice sound "mechanical"; thus, making the voice distinguishable from crew and communications voices.

An interesting point is that naturalness and intelligibility are not as strongly correlated with each other as they are with speech rate (Simpson et al., 1985). This is because a pronounced word may be intelligible even if it is not a natural pronunciation of the word. Speech rate is measured as the number of words spoken per minute, and has a significant effect on both intelligibility and naturalness (Simpson and Marchionda-Frost, 1984; Slowiaczek and Nusbaum, 1985). Merva (1987) reported that a speech rate of 180 words per minute is more intelligible than a speech rate of 210 words per minute for transcription tasks.

An important consideration for selecting a speech synthesizer is the required vocabulary size. Rule-base synthesizers have an unlimited vocabulary because the algorithm-based program pronounces words directly from text. Digitized speech is limited to a fixed vocabulary of either spoken words that are recorded or samples of digitally recorded waveforms. However, if the intended application is simply to

speak a small, fixed number of messages, then vocabulary size may not be an important consideration.

Research directed by David Pisoni and his colleagues at the Indiana University indicates that all speech synthesizers are not the same. Generally, digitized speech synthesizers are more intelligible than rule-based synthesizers. However, rule-based synthesizers are more versatile than digitized speech synthesizers, making them more attractive for a information system where the database will be updated periodically. Among the rule-based synthesizers, Digital Equipment Corporation DECtalk has been found to be significantly more intelligible than any of the other commercially available rule-based synthesizers (Manous et al., 1984).

DECtalk, version 2.0, was used in the screening study. DECtalk is a rule-based speech synthesizer which uses text-to-speech technology. This means that the computer analyzes each word as it is typed and applies rules of phonemes to generate the pronunciation of the word. A text-to-speech, rule-based speech synthesizer was selected because it offers the greatest flexibility due to its unlimited vocabulary capability. Because the department store information system was expected to have wide variety of store items, it was determined that the synthesizer must have the capability to speak a large number of words with a high level of intelligibility. DECtalk was selected because it meets this criterion and enjoys a general acceptance in the field as a high quality, premium speech synthesizer (Manous et al., 1984).

Application Issues for speech synthesis displays. More and more consumer products are implementing synthesized speech displays, with applications ranging

from informative displays in automobiles to novelty options on sewing machines. Consumer acceptance of such applications has been relatively low; the main reason being that the displays were unnecessary and/or unintelligible. However, there has been some success with implementing speech synthesis in information systems. User acceptance has been high for synthetic speech used in information retrieval, electronic mail, and information systems (Anderson, 1984; Gould and Boies, 1984; Schmandt, 1985a).

The reason for this contrast in successful versus ill-fated applications lies in the basic principles of systems design. Namely, before selecting speech synthesis as a display modality for an application, the designer must determine if the unique characteristics of speech synthesis match the display requirements for the user. In addition, synthesized speech must be compatible with the intended operating environment (e.g., noisy workplaces). Therefore, applying speech synthesis involves: (1) understanding user requirements, (2) knowing the capabilities of synthetic speech, (3) matching the display requirements for the application, and (4) satisfying environmental factors.

Synthetic speech is still a relatively new and developing technology. In the near future, we can expect speech synthesizers that approximate human speech and a plethora of new and innovative applications. Developing successful applications is dependent upon properly investigating the many ways synthetic speech affects user and system performance.

Telephone Information Systems

Telephone information systems are very popular in the consumer marketplace, but are rather scarce in the research literature. The main reason for this void is that many of these systems are developed by telephone companies, and much of the research is proprietary. In the available literature, telephone information systems which use synthetic speech were developed with a minimum amount of research (Hise and Lundin, 1985; Kidd, 1982; Rosson and Mellen, 1985; Waterworth and Lo, 1984) and often as case studies with no empirical research (Anderson, 1984; Gould and Boies, 1984; Podgorny, 1985; Schmandt, 1985a; Thomas, Rosson, and Chodorow, 1984; Witten and Mandams, 1977).

A literature review was performed as part of the engineering analysis for the research project. A total of 35 articles were identified as having relevant information pertaining to the design and evaluation of a telephone information system which uses synthetic speech. From the literature review, there exist several references for the 16 variables to be investigated. In the following paragraphs, the known effects of each of the variables are reviewed briefly. Because this study is designed specifically to investigate telephone information systems which use synthetic speech, the literature cited for each variable is limited to publications directly related to the subject area.

Type of voice. The type of voice and speech rate of the speaker are characteristics of the voice structure of synthetic speech. In previous research involving the DECtalk, version 1.8, the male voice Perfect Paul was found to be more intelligible and preferred over the female voice Beautiful Betty (Green, Manous and Pisoni, 1984). However, this research was conducted as a standard

intelligibility test. Thus, there is the unresolved question as to what voice is more intelligible and preferred over the limited bandwidth telephone communication lines. DECtalk Perfect Paul and Beautiful Betty, version 2.0, were tested in the screening study.

Speech Rate. Merva (1987) analyzed speech rate as part of her research with DECtalk, version 2.0. Merva found that subjects transcribed messages better at a speech rate of 180 words per minute (wpm) than at either 150 or 210 wpm. An interesting note in Merva's research is that transcription accuracy for the 210 wpm group improved over time and at the end of the experiment was approaching performance levels equal to that of the 180 wpm group. This suggests a possibility that experienced users can use systems with higher speech rates. In this study, speech rates of 180 and 240 wpm were tested.

Database organization. Database organization describes the structure of the menu hierarchies used to access the information in the system. Menu hierarchies are often used for telephone information systems (Kidd, 1982; Podgorny, 1985; Witten and Mandams, 1977) and are described in terms of their breadth (menu length) and depth (number of menu levels). The issue of menu breadth versus depth is a well investigated issue in visual display systems. The general finding is that longer menus with fewer levels are preferable to shorter menus with more levels (Kiger, 1984). However, little is known on this issue for auditory information systems.

Kidd (1982) conducted research on menu length and menu item position for auditory menu systems. Kidd reported no effect due to menu item position and a degradation in performance only when menu length exceeded 10 items. Kidd's

study was focused on addressing the issue of user problem solving when deciding on the correct menu option in an auditory menu. The study determined that auditory menus should be a simple design such that menu items are exclusive from one another. However, the study did not specifically consider the overall issue of the interrelationship between menu length and breadth, and how it affects the database organization. Menu length and depth are inversely proportional (i.e. as menu length increases, menu depth decreases, and vice versa).

For a telephone information system which requires users to listen to a list of menu items, menu breadth and depth probably has some effect on user performance given the results of the experiments referenced above. To evaluate database organizations, two hierarchies were evaluated in the screening study. One database was a 2x6 hierarchy, where there are 2 menu items per menu with 6 levels of menus. The other database was a 8x2 hierarchy, where there are 8 menu items with 2 levels per menus. Both designs have an equal number of information nodes (64) but have opposing organizations.

Number of target items. The number of items a user searches for should be considered in the design of a telephone information systems. The issue is that a person searching for more than one item during a single phone call to the system will begin each subsequent search from the previous store item node rather than the main menu. Therefore, if users are expected to perform multiple searches in the system, then the telephone system should be designed to facilitate multiple searches. This issue has not been addressed in the reviewed literature, but was identified as an important design issue in the engineering analysis. In the screening study, single item and double item target searches were investigated.

System response time. The time it takes for a computer to respond to a user input is defined here as system response time. System response time was a major concern in early computing systems using visual systems, because users usually had to wait a relatively substantial amount of time for the computer to perform even the simplest of tasks (Engel and Granda, 1975). This is no longer the case with modern day computing systems. Most systems respond almost instantaneously to what used to be complex and time consuming procedures. The issue of system response time in an telephone information system is really a question of pacing or governing the response time of the speech display. If the systems responds too quickly, the speech synthesizer might begin speaking before the user is ready to listen. Thus far, this issue has not been reported in the literature, but was identified during the engineering analysis as having the potential of being an important factor.

Input timeout. Input timeout is the amount of time the information system pauses between menu items to allow a user to select the menu item just spoken. The longer the input timeout, the more time a user has to select the previously spoken menu item. A longer input timeout increases the overall time it takes to locate a store item. A shorter input timeout, reduces the amount of time a user has to select the spoken menu item and increases the likelihood that the user will not make a selection within the required time limit. In previous information systems, input timeout has been designated by the designer without any reported rationale (Anderson, 1984; Gould and Boies, 1984; Hise and Lundin, 1985; Kidd, 1982; Podgorny, 1985; Schmandt, 1985a; and Witten and Mandams, 1977). The screening study started this investigation by considering the shortest possible input timeout of 2 seconds and a longer input timeout of 4 seconds. The resident

computer software has a built in variance of ± .50 seconds for the input timeout routine, making 2 seconds the shortest acceptable time value. Four seconds was identified as a sufficiently long input timeout during system prototyping and pretesting. Prototype testing indicated that subjects typically made a menu selection within 4 seconds of hearing the keyword.

Pause/resume speech. An inherent feature of a visual information system is that the user can pause and resume the task at his/her convenience. Because of the transient nature of speech, pause/resume is not an inherent feature of a menu driven, auditory information system. Auditory systems proceed with a new action with every user input. Therefore, a user command option of pause/resume speech must be designed into the system. Such a feature provides the user with an option to pause anytime during a search task when he/she feels a break or "timeout" is necessary. Schmandt (1985a) and Witten and Madams (1977) designed this feature into their respective telephone information systems. Both systems were reported as case studies and do not provide any information as to the utility of the pause/resume feature. The engineering analysis also identified this feature as being an important feature that would enhance the usability of the system. The screening study investigated the pause/resume command as a telephone keypad input. The command was provided to some subjects and not provided to others. The command was only valid during the searching portion of the experimental task. Subjects were not able to pause the system when they were at the store item or during the transcription task.

Repeat keyword. A repeat keyword command option is an important feature in auditory displays. If the spoken keyword is masked by some extraneous

noise, the keyword can be heard incorrectly or missed altogether. In addition, the limitations of rule-based synthetic speech may cause certain words to be hard to understand, causing the user to want to hear the word again. A repeat keyword feature provides the capability to have keywords repeated at the listener's request. A repeat speech command option is a popular design feature in existing telephone information systems (Kidd, 1982; Rosson and Mellen, 1985; Schmandt, 1985a; and Witten and Mandams, 1977). Merva (1987) also investigated the effects of message repetition and found that message repetition significantly improved message transcription accuracy. In the screening study, the repeat keyword command was provided as a command option on the telephone keypad, similar to that for pause/resume speech. The command is only valid during the searching portion of the experimental task. Subjects were not able to repeat information messages when they were at the information node. This restriction existed to protect the integrity of measuring the effects of speech rate and voice type on transcription accuracy.

Spell-out keyword. A user could select spell-out keyword whenever a keyword was not understood. While the keywords in the database were selected because they are generally intelligible, a spell-out capability is important for words that might be unintelligible to certain users. Rosson and Mellen (1985) and Schmandt (1985a) both provided this capability to system users, but did not report on its value or utility. This screening study investigated the value and utility of the spell-out speech command option. Users were able to have a keyword spelled-out for them by pressing an assigned key on the telephone keypad after the keyword was spoken. Subjects were not able to spell-out information messages when they

were at the store item. This restriction existed to protect the integrity in measuring the effects of speech rate and voice type on transcription accuracy. Half of the subjects had the capability to have keywords spelled-out.

Wallet guide. User aids are very helpful for system users with no training or little experience. In this screening study, naive users were tested with various types of user aids. Wallet guides were provided to half of subjects to assist them in using the information system. Previous research has used telephone keypad overlays or command templates to remind users of the function of each telephone key (Podgorny, 1985; Witten and Madams, 1977). Such guides basically listed out the functions of each telephone key, but did not provide graphical representations of the database system. To investigate the value of a visual representation of the system in helping users find the desired item in the system, a wallet guide which graphically represents the database hierarchy was provided to half of the subjects. The introduction of a visual representation of an auditory system introduced a second display modality with unknown effects on user performance.

Menu selection, database, and command feedback. Feedback in the form of input echoing is another type of aid to the user. The telephone by design provides a tone for input echoing. However, a tone provides little information as to what action a subject has just selected. Spoken input echoing provides a clear feedback message as to what action was just selected. Earlier studies and systems have used this aid to assist users in using information systems (Gould and Boies, 1983; Nakatani and O'Connor, 1980; and Podgorny, 1985). Surprisingly, Witten and Madams (1977) purposely eliminated key-selection feedback, arguing that the

subject would be aware of his/her actions by listening to what the system did next (implicit feedback).

In this study, three different types of feedback were provided: menu selection feedback, database movement feedback, or command selection feedback. Menu selection feedback preceded each new menu by stating "*Keyword* options are..." for the keyword that was just selected. Database movement feedback was provided when a subject selected to back-up one menu by stating "Previous menu options are...", or when a subject selected the main menu by stating "Main menu options are...". Finally, command selection feedback was provided when a subject selected either pause/resume, repeat keyword, or spell-out keyword. Command feedback was presented for the three command options as an introductory message stating which command had been selected. For pause/resume the message was "Conversation on pause, press the 7 key to continue"; for repeat keyword, "Repeat selected *Keyword*" when the subject wanted *keyword* repeated; and for spell-out keyword, "Spell selected *K-E-Y-W-O-R-D*."

Background music. Environmental considerations include the effects of background noise, activities and operator workload on user performance. Background music is a common occurrence in locations where a telephone information system could be used (e.g., workplace, shopping malls, airports, homes). Thus, background music is an issue which should be considered for this auditory system. In all previous studies involving synthetic speech and/or information systems, there has been a significant degradation in performance due to noise (Logan, Pisoni, and Greene, 1985; Pisoni, 1979; Simpson and Marchionda-Frost, 1984). However, the literature review did not identify any previous

investigation of the effects of background music on user performance. Background music was played in this study for half of the subjects.

Subject age and sex. User variables such as age and sex usually do not have a significant effect on user performance in information systems. However, in telephone information systems the results are not clear. There exists a definite conflict in whether age has an effect on user performance (Rosson and Mellen, 1985; Waterworth and Lo, 1984). However, the general literature on aging is clear that hearing ability, especially in the higher frequencies, decreases with age (Sanders and McCormick, 1987). Sex of the listener is an unresolved issue, because it typically was not manipulated in the reviewed literature on synthetic speech. In this study, both sexes were tested as well as the college age group (ages 18 to 30) and the middle age group (ages 45 to 60).

Screening Studies

Screening studies are a class of fractional factorial designs used to determine which of many factors have non-trivial effects on the performance of certain tasks (Simon, 1977). They are a valuable experimental tool to the researcher who must determine what variables out of a large identified set should be investigated in greater detail. However, they are intended to be used to identify important factors, and not to obtain an accurate representation of the experimental space for the variables under consideration (Simon, 1977).

Fractional factorial designs are used when the researcher is interested in so many variables that a complete factorial experiment is not feasible, or when conditions prohibit the researcher from running the proper number of trials for a

complete factorial experiment (Cochran and Cox, 1957). In a fractional factorial design, only a portion of the amount of variable combinations are investigated. By appropriately selecting which experimental conditions (i.e., combinations of variable levels) need to be investigated, significantly valid results can be obtained which are robust to experimental error. Algorithms or pre-designed matrices are used to determine what trial conditions should be tested. The desired number of variables, trials, subjects, and resolution level determine what fractional factorial design should be used.

The real worth of screening studies lies in their inherent ability to assess a large number of variables with only a relatively small number of observations. Simon and Roscoe (1984) used a screening study to evaluate six training equipment and user factors in research on transfer of training. The authors reported that the study was the first time six training factors were analyzed at once. Tatro and Roscoe (1986) also found the screening study to be worthwhile in their research on integrated flight displays. The authors were able to investigate the effects of eight display factors on pilot performance. Both articles stated that the use of fractional factorial designs allowed many more variables to be considered than if full factorial designs had been used. In addition, both papers also found that the fractional factorial design required far fewer data points, making their respective research very economical.

Whitehurst (1982) used a screening study to investigate the effects of eight display factors on the readability of moving-pointer, fixed scale dials. In this study, two different screening designs were used to increase the confidence that the independent variables, and not the confounded interactions, accounted for a

particular share of the total variance. Thus, if each screening design was properly constructed, then the percentage of variance accounted for by each variable should be approximately the same. The results showed that significant independent variables in each design accounted for approximately equal percentages of total variance and that three-way interactions accounted for a small amount of total variance.

Whitehurst also tested the unconfounded effects of black-on-white versus white-on-black dials by repeating the entire experiment with a separate set of subjects. Repeating the experiment for the sake of testing one more variable is in direct conflict with the philosophy of data efficiency in conducting screening studies, and Whitehurst (1982) did not offer any explanation for this methodology The results of the two experiments were similar with the additional finding that dial background-foreground contrasts had significant effects on readability.

Given such a powerful capability, the question is rightly asked; "Why are screening studies so unpopular?" Researchers have traditionally focused their attention towards specific cause-effect studies. Such research strategies typically concentrate on a well bounded, but very small area in the total domain of the research topic. As a result, knowledge of the total domain grows slowly and is often incomplete. Numerous in-depth studies are conducted which determine both trivial and significant results. While this knowledge fills in the gaps of uncertainty, the economics of obtaining that knowledge is questionable.

The argument for screening studies is secondary to the argument for sequential research strategies -- an overall methodological approach for researching a topic. Screening studies are but one tool used in sequential research strategies.

The role of screening studies is to expedite the analysis of what variables have noticeable (significant) effects and what variables have trivial (insignificant) effects (Simon, 1977). Those variables identified as having significant effects can then be investigated in greater detail using more appropriate techniques, such as complete factorial designs.

Sequential research strategies provide planned direction for researching a behavioral topic. By identifying all possible variables and factors which might affect user performance, a boundary for the topic is set. From within this perimeter, literature reviews and engineering analyses can be conducted to identify which variables out of the total set are most appropriate for a screening study. The screening study provides an efficient means of assessing this first set of variables and establishing preliminary findings on their effects on user performance. Later studies can be planned based on what was learned in the screening studies. The end result is a series of experiments which can be related to one another and which are planned based on a systematic investigation of all variables.

METHOD

The screening study investigated how 16 independent variables affect 23 dependent measures of operator performance. A Hadamard 32x32 matrix design was used to investigate the independent variables (Diamond, 1981). Hadamard matrices are well suited for screening purposes, where the investigator is interested in the main effects of a large number of independent variables. The design requires only 32 trials, whereas a full factorial design would require 65 536 trials.

Experimental Design

The assignment of variable levels was obtained by constructing a Hadamard 32x32 matrix and assigning a variable to pre-determined contrast columns (Diamond, 1981), see Table 2. Each letter above a contrast column represents one variable; variable assignments are listed in Table 3. The numbered contrast columns represent the pre-determined alias sets of two-way interactions. In the matrix, high levels of a variable are indicated by a plus sign (+), while low levels are indicated by a minus sign (-). The matrix is a balanced design, robust to trend effects.

For this application, the Hadamard 32x32 matrix design confounds main effects with three-way and higher order interactions, and two-way interactions with other two-way interactions. By assuming that three-way and higher order interactions are non-significant, main effects can be determined. Because two-way interactions are aliases with one another, specific two-way interactions cannot be investigated. Table 4 provides a summary of the alias structure for the 16 independent variables in the screening study.

Table 2. Screening Study 32x32 Hadamard Matrix

Trial (Subject)	Contrasts																															Hi-Level Treatments
	A	B	C	D	E			F	G		H	I	J	K	L			M		N		O	P									
1	+	+	-	-	-	+	+	-	+	+	-	+	-	+	-	-	-	-	+	-	-	+	-	+	+	-	-	+	+	+	+	aefghjop
2	+	+	+	-	-	-	+	+	-	+	+	-	+	-	+	-	-	-	-	+	-	+	+	-	-	+	+	+				abfgimnp
3	+	+	+	+	-	-	-	+	+	-	+	+	-	+	-	+	-	-	-	-	+	-	+	-	+	+	-	-	+			abcghjkn
4	+	+	+	+	+	-	-	-	+	+	-	+	+	-	+	-	+	-	-	-	-	+	-	+	-	+	+	-	-	+		abcdfhio
5	+	+	+	+	+	+	-	-	-	+	+	-	+	+	-	+	-	+	-	-	-	-	+	-	-	+	-	+	+	-	-	abcdefghijklmnop
6	+	-	+	+	+	+	+	-	-	-	+	+	-	+	+	-	+	-	+	-	-	-	-	+	-	-	+	-	+	+	-	bcdegijp
7	+	-	-	+	+	+	+	+	-	-	-	+	+	-	+	+	-	+	-	+	-	-	-	-	+	-	-	+	-	+	+	cdehjklo
8	+	+	-	-	+	+	+	+	+	-	-	-	+	+	-	+	+	-	+	-	+	-	-	-	-	+	-	-	+	-	+	adehiknp
9	+	+	+	-	-	+	+	+	+	+	-	-	-	+	+	-	+	+	-	+	-	+	-	-	-	-	+	-	-	+	-	abefijkl
10	+	-	+	+	-	-	+	+	+	+	-	-	+	-	+	+	-	+	-	+	+	-	-	-	-	-	+	-	-	-	+	bcfgjlmo
11	+	+	-	+	+	-	-	+	+	+	+	-	-	+	-	+	+	-	+	-	+	-	+	-	-	-	-	+	-	-	-	acdfgklp
12	+	-	+	-	+	+	-	-	+	+	+	+	-	-	+	-	+	+	-	+	-	+	-	+	-	-	-	-	+	-	-	bdefghkm
13	+	-	-	+	-	+	+	-	-	+	+	+	+	-	-	+	-	+	+	-	+	-	+	-	+	-	-	-	-	+		cefghiln
14	+	+	-	-	+	-	+	+	-	-	+	+	+	+	-	-	+	-	+	+	-	+	-	+	-	+	-	-	-	-		adghijlm
15	+	-	+	-	-	+	-	+	+	-	+	+	+	+	-	-	+	-	+	+	+	-	+	-	+	-	-	-				behijmno
16	+	-	-	+	-	-	+	-	+	+	-	+	+	+	+	-	-	+	-	+	+	+	-	+	-	+	-	-				cfhijkmp
17	+	-	-	-	+	-	-	+	-	+	+	-	+	+	+	+	-	-	+	-	+	+	+	-	+	-	+	-				dfgijkno
18	+	-	-	-	-	+	-	-	+	-	+	+	-	+	+	+	+	-	-	+	-	+	+	+	-	+	-	+				egjklmnp
19	+	+	-	-	-	-	+	-	-	+	-	+	+	-	+	+	+	+	-	-	+	-	+	+	+	-	+	-				afhklmno
20	+	-	+	-	-	-	-	+	-	-	+	-	+	+	-	+	+	+	+	-	-	+	-	+	+	+	-	+				bghiklop
21	+	+	-	+	-	-	-	-	+	-	-	+	-	+	-	+	+	+	+	+	-	-	+	-	+	+	+	-				acijlnop
22	+	-	+	-	+	-	-	-	-	+	-	+	-	+	+	-	+	+	+	+	+	-	-	+	-	+	+	+				bdfhjlnp
23	+	+	-	+	-	+	-	-	-	-	+	-	+	-	+	+	-	+	+	+	+	-	-	+	-	+	+					acegikmo
24	+	+	+	-	+	-	+	-	-	-	-	+	-	+	-	+	+	-	+	+	+	+	-	-	+	-	+					abdjkmop
25	+	+	+	+	-	+	-	+	-	-	-	-	+	-	+	-	+	+	-	+	+	+	+	-	-	+	+	-				abcehlmp
26	+	-	+	+	+	-	+	-	+	-	-	-	-	+	-	+	-	+	+	-	+	+	+	+	-	-	+	+				bcdiklmn
27	+	+	-	+	+	+	-	+	-	+	-	-	-	-	+	-	+	-	+	+	-	+	+	+	+	-	-	+				acdefjmn
28	+	+	+	-	+	+	+	-	+	-	+	-	-	-	-	+	-	+	-	+	+	-	+	+	+	+	-	-				abdeglno
29	+	-	+	+	-	+	+	+	-	+	-	+	-	-	-	-	+	-	+	-	+	+	-	+	+	+	+	-				bcefknop
30	+	-	-	+	+	-	+	+	+	-	+	-	+	-	-	-	-	+	-	+	-	+	+	-	+	+	+	-				cdghmnop
31	+	-	-	-	+	+	-	+	+	+	-	+	-	+	-	-	-	+	-	+	-	+	+	-	-	+	+	+				defilmop
32	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-				(1)

Column Assignments:

0 A B C D E 6 7 8 F 6 11 H I J 15 K 17 L 19 20 21 22 M 24 25 N 27 0 P 30 31

Table 3. Coding Assignment for the Independent Variables

A = Type of Voice
B = Speech Rate
C = Database Organization
D = Number of Targets Items
E = System Response Time
F = Input Timeout
G = Pause/Resume Speech
H = Repeat Speech
I = Spell-out Speech
J = Wallet Guide
K = Menu Selection Feedback
L = Database Movement Feedback
M = Command Selection Feedback
N = Background Music
O = Age of Subject
P = Sex of Subject

Table 4. Alias Structure for the 16 Independent Variables

Contrast	Aliases
0	Not Defined
1	A(*****)
2	B(*****)
3	C(*****)
4	D(*****)
5	E(*****)
6	AC,DF,BH,GK,EM,JN,IO,LP
7	BD,EG,FH,CI,KM,LN,AO,JP
8	CE,GI,DJ,HL,AM,FN,KO,BP
9	F(*****)
10	G(*****)
11	AE,FJ,IK,BL,CM,DN,GO,HP
12	H(*****)
13	I(*****)
14	J(*****)
15	EF,GH,AJ,BK,IL,DM,CN,OP
16	K(*****)
17	AG,HJ,CK,DL,IM,BN,EO,FP
18	L(*****)
19	AB,CH,FI,JK,EL,GN,DO,MP
20	BC,AH,DI,GJ,LM,KN,FO,EP
21	CD,AF,BI,EJ,KL,MN,HO,GP
22	DE,BG,CJ,HK,FM,AN,LO,IP
23	M(*****)
24	BF,DH,AI,EK,JL,GM,CO,NP
25	CG,EI,BJ,AK,FL,HN,MO,DP
26	N(*****)
27	FG,EH,DK,CL,BM,IN,JO,AP
28	O(*****)
29	P(*****)
30	AD,CF,HI,GL,JM,EN,BO,KP
31	BE,DG,IJ,FK,AL,HM,NO,CP

Adapted from: Diamond, W.J. (1981). *Practical experiment designs for engineers and scientists*. Belmont, CA: Lifetime Learning.

***** Third-order and higher interactions are aliases with main effects.

The study was conducted as a between-subject experiment. While not the most efficient design, a between-subject design was the only acceptable design. The two user variables weretween-subject variables by default (age, sex), therefore eliminating the possibility of a within-subject design. Furthermore, the design matrix prohibited a mixed factor design, because trial conditions could not be balanced with subjects. There was also the strong possibility that on-line training would greatly affect subject performance across trial conditions. Thus, the screening study was conducted as a between-subject experiment with thirty-two subjects tested, one subject per trial condition. With only one subject per trial condition, the effect due to subject error was pooled with all other sources of experimental error. Each subject had a unique combination of variables (Table 5).

Subjects

Thirty-two, native English speaking subjects volunteered to participate in this study. Eight middle age females, eight middle age males, eight college age females, and eight college age males comprise the test subjects who participated in this study. Subjects had no previous experience with synthesized speech and received compensation for their time. Subjects completed a pretest demographics questionnaire (see Appendix II). Subjects were also be given a hearing test to ensure they had acceptable hearing levels in both ears. The test consisted of subjects acknowledging when they heard a 3 second pulsed tone presented to either their right or left ear. Tones were presented at 26 dB(A) at the following frequencies: 750, 1000, 2000, and 4000 Hz. If a subject failed to acknowledge the tones at a given frequency, the subject was presented with the tones at that

Table 5. Assignment of Treatment Conditions to Subjects.

Subject	Voice Type	Speech Rate	Dbase Org	No. Targets	Sys Resp	Input Timeout	Pause Resume	Repeat Speech	Spell-out	Wallet Guide	Select. F-back	Menu F-back	Cmnd F-Back	Music	Age	Sex
1	Paul	180	8x2	1	4	4	A	A	NA	A	NA	NA	NA	office	MA	F
2	Paul	240	8x2	1	0	4	A	NA	A	NA	NA	NA	A	mall	CA	F
3	Paul	240	2x6	1	0	2	A	A	NA	A	A	NA	NA	mall	CA	M
4	Paul	240	2x6	2	0	4	NA	A	A	NA	NA	NA	NA	office	MA	M
5	Paul	240	2x6	2	4	4	A	A	A	A	A	A	A	mall	MA	F
6	Betty	240	2x6	2	4	2	A	NA	A	A	NA	NA	NA	office	CA	F
7	Betty	180	2x6	2	4	2	NA	A	NA	A	A	A	NA	office	MA	M
8	Paul	180	8x2	2	4	2	NA	A	A	NA	A	NA	NA	mall	CA	F
9	Paul	240	8x2	1	4	4	NA	NA	A	A	A	A	NA	office	CA	M
10	Betty	240	2x6	1	0	4	A	NA	NA	A	NA	A	A	office	MA	M
11	Paul	180	2x6	2	0	4	A	NA	NA	NA	A	A	NA	office	CA	F
12	Betty	240	8x2	2	4	4	A	A	NA	NA	A	NA	A	office	CA	M
13	Betty	180	2x6	1	4	4	A	A	A	NA	NA	A	NA	mall	CA	M
14	Paul	180	8x2	2	0	2	A	A	A	A	NA	A	A	office	CA	M
15	Betty	240	8x2	1	4	2	NA	A	A	A	NA	NA	A	mall	MA	M
16	Betty	180	2x6	1	0	4	NA	A	A	A	A	NA	A	office	CA	F
17	Betty	180	8x2	2	0	4	A	NA	A	A	A	NA	NA	mall	MA	M
18	Betty	180	8x2	1	4	2	A	NA	NA	A	A	A	A	mall	CA	F
19	Paul	180	8x2	1	0	4	NA	A	NA	NA	A	A	A	mall	MA	M
20	Betty	240	8x2	1	0	2	A	A	A	NA	A	A	NA	office	MA	F
21	Paul	180	2x6	1	0	2	NA	NA	A	A	NA	A	NA	mall	MA	F
22	Betty	240	8x2	2	0	4	NA	A	NA	A	NA	A	NA	mall	CA	F
23	Paul	180	2x6	1	4	2	A	NA	A	NA	A	NA	A	office	MA	M
24	Paul	240	8x2	2	0	2	NA	NA	NA	A	A	NA	A	office	MA	F
25	Paul	240	2x6	1	4	2	NA	A	NA	NA	NA	A	A	office	CA	F
26	Betty	240	2x6	2	0	2	NA	NA	A	NA	A	A	A	mall	CA	M
27	Paul	180	2x6	2	4	4	NA	NA	NA	A	NA	NA	A	mall	CA	M
28	Paul	240	8x2	2	4	2	A	NA	NA	NA	NA	A	NA	mall	MA	M
29	Betty	240	2x6	1	4	4	NA	NA	NA	NA	A	NA	NA	mall	MA	F
30	Betty	180	2x6	2	0	2	A	A	NA	NA	NA	NA	A	mall	MA	F
31	Betty	180	8x2	2	4	4	NA	NA	A	NA	NA	A	A	office	MA	F
32	Betty	180	8x2	1	0	2	NA	NA	NA	NA	NA	NA	NA	office	CA	M

A = Available MA = Middle Age F = Female
NA = Not Available CA = College Age M= Male

frequency and decibel level two more times. For retesting on any frequency, subjects were required to acknowledge correctly the tones both times in order to pass the hearing test at that frequency. Subjects not able to pass the hearing test still participated in the experiment, but their data were not included in this study. Their data were retained as happenstance data for possible future studies involving hearing impaired users.

To help control for homogeneity in subject variance a time limit criterion was imposed on store item search time. If a subject failed to find any store item, either a practice or an experimental store item, within 15 minutes of starting the search, then that subject's data were automatically disqualified, and the treatment condition was retested with a new subject. Subjects were not made aware of this test criterion, so as not to impose any time-related anxiety. Fifteen minutes was selected because it represented an excessive amount of time to find a store item based on experimental pretesting.

Materials

Speaker Telephone. A speaker telephone was used so that subjects did not have to hold a handset to their ear during the experiment. The speaker telephone was a Panasonic Easaphone. The telephone speaker had a volume adjustment control which was fixed at a measured level of 73.8 dB(A) to prevent subjects from adjusting the volume. General background noise was measured at 60.6 dB(A).

Computer. A DEC VAX 11/750 executed the application software. A DEC VT220 was used to present the 16 targets and record the information message transcriptions and subjective ratings.

Stereo System. A GE stereo VCR and stereo TV were used to present the video instructions and background music. The background music was a 120 minute recording of the album Muzak of the '80s, by Muzak Company. The music was played at a measured level (integrated sound level) of approximately 64 $\pm$ 2 dB(A) and was located directly behind the subjects at a distance of 6 feet.

Speech Synthesizer. The speech synthesizer was a Digital Equipment Corporation DECtalk, Version 2.0. Voice types tested were Perfect Paul and Beautiful Betty. The DECtalk unit was selected based on its performance in intelligibility studies conducted at Indiana University. Green, Manous and Pisoni (1984) found the DECtalk, version 1.8 to be the most intelligible speech synthesizer of four commercially available speech synthesizers. Green and his colleagues tested the DECtalk, version 1.8; the Prose-2000, version #8-84; the MITalk-79; and the Type-n-Talk, version 3-82. DECtalk's Perfect Paul was found to be more intelligible than any other DECtalk voice, and far more intelligible than any other synthesizer. DECtalk's Beautiful Betty was the second most intelligible voice in the study.

Manual phoneme or stress coding was not used to improve or enhance pronunciation by the DECtalk voices. Keywords and information messages were entered into the database as they are normally spelled and spoken using the internal text-to-speech conversion algorithms. The only exception is that compound words were entered with hyphens at the appropriate location to reduce mispronunciation (e.g., basket-ball, sweat-pants).

Information Databases. Two information databases were constructed for the research project (see Appendix IV). Symmetrical databases containing the same

number of terminal (bottom level) nodes were selected to permit comparison of alternative database organizations. One database was a 2x6 hierarchy, where there were 6 levels of menus with each menu having 2 items. The second database was an 8x2 hierarchy, where there were 2 levels of menus with each menu having 8 items.

Both databases had the same 64 store items (terminal nodes) and information messages. The differences between the two databases, other than the different hierarchies, were the menu items (keywords). A keyword is a title for a group of related items (e.g. "automotive" is a keyword for "tires" and "motor oil"). In the 2x6 hierarchy, keywords were brief and usually a single word (e.g. dresses). In the 8x2 hierarchy, keywords were usually two- and sometimes three-word noun clauses (e.g. women's dresses). The 8x2 hierarchy had far fewer menu items (72) than the 2x6 hierarchy (126), and therefore required more information for each menu item.

Keywords and Information Messages. Keywords were developed by back-fitting a department store into the database hierarchies. The criteria for selecting keywords were that the bottom level keywords would be a specific store item (refrigerators) or type of store item (pearl earrings). Because of the 2x6 database, only store items which could be grouped into sets of 2, 4, 8, 16, 32, and 64 could be used. The 8x2 database was constructed by regrouping store items sets from the 2x6 database. In addition, a concerted effort was made to use only sets of store items which were fairly distinct from other sets of store items. This was done to reduce searching errors due to semantics or ambiguous keywords.

Information messages were of four types: Location, Price, Availability, or Information. Each message had the form of *Modifier subject verb preposition modifier object* (i.e. "Leather coverings are guaranteed for five years"). A large set of information messages were constructed with each modifier, subject, and object used only once. As an example, the word "leather" could only be in one information message. The verbs and prepositions used for each message are shown in Table 6. By standardizing the message format and using a small set of verbs and prepositions, the middle part of each information message became familiar to the subject and did not provide clues as to the meaning of the message. The first two words and last two words in an information message were scored for transcription accuracy (Merva, 1987).

Keywords and information messages were pretested to determine if any were obviously unintelligible or misunderstood. Keywords were tested using DECtalk voice settings of Perfect Paul at 180 wpm. These voice settings were used because they are generally accepted as the most intelligible voice for the DECtalk speech synthesizer (Green et al., 1984; Merva, 1987). Certain keywords and store items were found to be unintelligible and were replaced with synonyms or similar items. This action should have reduced unwanted errors due to poor word selection, but should not have biased the results of the study. Keywords for each database are listed in Appendix V.

Information messages were pretested in a small study involving subjects listening and transcribing information messages. Seventy-seven information messages were developed for the four types of information messages: 19 location messages, 19 price messages, 20 availability messages, and 19 information

Table 6. Format of the Information Messages

Information Type	Format
LOCATION:	*Modifier Subject* is/are in *modifier object.* on near
PRICE:	*Modifier Subject* is/are reduced by *modifier object.* for
	Modifier Subject are sold by *modifier object.* for
AVAILABILITY:	*Modifier Subject* is/are available at *modifier object.* by in with
INFORMATION:	*Modifier Subject* are offered on *modifier object.* to for with
	Modifier Subject is/are required on *modifier object.* to for within

messages. The messages were tested under two voice conditions. As with the keyword testing Perfect Paul at 180 wpm was used as a baseline voice setting for testing the messages. Betty at 240 wpm served as an extreme alternative voice setting in order to test whether or not voice settings affect intelligibility.

Six college age students served as subjects in the pretest. Each was given and passed the same hearing test as described above. Three subjects were assigned to each voice setting and were presented with written instructions and listened to spoken instructions from the DECtalk voice assigned for each subject. Both the spoken instructions and the information messages were presented through the speaker telephone. Subjects were given 5 practice messages to familiarize themselves with the task and the voice setting. The practice messages were randomly selected Harvard Psychoacoustic Sentences (Allen et al., 1987).

The experimental task was subject-driven, requiring subjects to press the spacebar key on a computer keyboard to hear the next message. The subject was instructed to listen carefully to the entire message and then transcribe the message onto a sheet of paper. Subjects were also asked to underline any words in a message that they were uncertain of in their transcription.A total of 82 information messages was transcribed by each subject. Sentences were scored on correctly transcribing the adjective for the subject, the subject, the adjective for the object, and the object for each message.

The purpose of the pretesting was two-fold. First, to identify sentences which were reasonably intelligible but not excessively easy or hard. The test criterion was that at least one word in an information message would be missed by at least one subject for each voice setting, but that a word in an information message

was not missed by all three subjects for each voice setting. The objective was to select messages which were susceptible to transcription errors but not unintelligible or completely intelligible regardless of voice setting. Second, to determine if the extreme voice setting (Betty at 240 wpm) had any noticeable effect on intelligibility as compared to the baseline voice setting (Paul at 180 wpm). The objective was to determine if the intelligibility of information messages was affected by alternative voice characteristics. If no noticeable effect was determined, then the information transcription task would not be a valid, desirable experimental task.

The results of the pretesting identified 42 potential information messages for the screening study. From this set, 32 information messages (8 messages for each information type) were selected. The 32 information messages are listed with their associated keyword in Appendix V. The analysis also found a significant ($p < .05$) difference in transcription accuracy between the two voice settings, with Paul at 180 wpm having a higher transcription accuracy score than Betty at 240 wpm.

Both databases were also pretested to validate the design of the menus. In the pretest, 3 trials for each database were run, with each run having random assignments of levels for the remaining 15 variables. Six college age subjects participated in the experiment, with each subject assigned to 1 trial condition. Both databases were usable by the subjects, and none of the other 15 variables appeared to have a detrimental effect on user performance.

Pretest results also clearly indicated that subjects who were not given written and verbal instructions did not fully understand how the system worked. For the first 2 to 4 searches, these subjects had the system hang up on them when they failed to select a menu item. However, once the reason for the system hanging

up was explained to these subjects, all of the subjects began to locate target items adequately. One subject, given a demonstration of the target search task in addition to written and verbal instructions, had no difficulty performing any of the target searches. This result strongly suggests the need for a demonstration of the system and task as part of the instructions to the subject.

As a result of the pretesting several changes were made to the experimental task. First, the system was modified such that the system did not hang up on a subject after a menu is spoken twice. This change was made to eliminate the possibility of lost data points for a target. Second, the system only spoke the information message when the correct store item was located. If the subject located an incorrect store item, the system stated the store item and informed the subject to continue searching. Third, the pause/resume, repeat, and spell-out command options were only active during the target search and were not active at the information message. This was done to prevent biasing the data for the transcription task which is designed to measure the effects of voice characteristics on transcription accuracy. Fourth, a videotape of of how the system works and a demonstration of how the target search task is performed was made. The videotape was played for each subject during the instructional part of the experiment. Finally, each subject was given two practice targets after the instructions were completed. The practice targets were added to familiarize the subjects with the experimental task and reduce the effects of learning identified in the database pretesting.

Procedure

The experimental procedure followed the outline of activities listed in Table 5. At the beginning of the experimental session, subjects were first given an informed consent form to read and sign (see Appendix I). If the subject consented to participate in the experiment, the subject completed a pretest questionnaire for demographic purposes (see Appendix II). The subject was then given a hearing test, as described earlier, to ensure that the subject had sufficient hearing. The experimenter maintained a diary for each subject to record any significant events during the experiment or comments a subject might make.

The instructions began with an introduction to the information system and how it worked, followed by a detailed set of instructions (see Appendix III). The subject read the introduction and the instructions while the DECtalk (set at the assigned voice settings) spoke the introduction and instructions. Next, the subject watched a videotape which repeated the instructions and demonstrated how the task should be performed. Finally, the subject was presented with oral (DECtalk) and written instructions clarifying that the system in the video may be different than the actual system in some ways, and also stated what functions were available on the telephone keypad. The experimenter was present to answer any questions on how to perform the experimental task. To ensure that the subject understood the task, the experimenter asked the subject to recap the steps of the task.

The subject was then instructed to begin the practice target searches by calling the information system. After the practice targets were completed, any questions the subject might have were answered and the subject was instructed to begin the experiment by calling the information system.

The experimental task involved the following: subjects searched for specific store items in a telephone information system for a fictitious department store (Hokie Wholesale). Subjects called the information system on the speaker telephone to find information on selected store items. Store items were presented as targets on a computer display terminal in front of the subject. The target requested that the subject find the information message for a specific store item (or two store items for multiple searches). For a single store item search, the target message would read "What is the information message for golf books?" For a multiple store items search, the target message would read "What are the information messages for hand cream and golf books?" The target message remained displayed during the entire target search. Fifteen seconds after the target message was displayed a "ready..." message was displayed on the computer display screen to indicate that the search about to begin. Two seconds after the ready message, a "begin the search for the store item" message was displayed on the screen, and the information system spoke the first keyword in main level menu.

The information system worked by speaking menus of keywords. A keyword is a title for a group of related items (e.g., "automotive" is a keyword for a group of items like tires, and motor oil). When the subject heard a keyword which most closely relates to the target store item, the subject selected that keyword by pressing the "#" key on the telephone keypad. The system then spoke the first keyword in a new menu of keywords related to the previously selected keyword (e.g, if "automotive" was selected rather than "sporting goods", then the next menu spoken would be " tires ... motor oil"). By selecting the appropriate keywords, the subject located the store item in the information system.

Once the subject had located the target store item, the information system stated: "At store item *keyword*, press the 2 key to hear the information message." The information message had something to do with the price, location, availability, or information about the store item or information relevant to shopping in the department store. If the subject reached the wrong store item, the information system stated: "At store item *keyword*, continue searching." The subject would then have to leave that store item and continue searching for the correct store item.

Once the subject had located the store item, the subject was prompted to press the 2 key on the telephone keypad to hear the information message. The system spoke the information message and directed the subject to "Begin transcription". The computer screen then displayed a message requesting that the subject transcribe the information message just heard. There was no time limit for the transcription task and subjects were encouraged to be accurate rather than expedient in their transcription. Subjects were instructed to transcribe whatever they heard, even if they were not entirely sure of their answer.

After the transcription was entered, the computer screen display prompted subjects to rate how certain they were of their answer on a scale of 1 (very uncertain) to 7 (very certain). The scale was a bi-polar adjective scale. Similarly, subjects were then asked to rate how difficult it was to understand the message. Again, the rating scale was a bi-polar adjective scale from 1 (very difficult) to 7 (very easy). Finally, subjects rated the difficulty of searching for the store item in the database. Once again, the rating scale was from 1 (very difficult) to 7 (very easy). Examples of these subjective rating scales are found in Appendix VI.

In the case where a subject was performing a multiple search, the original target had two store items listed (e.g., "What are the information messages for hand cream and golf books?"). After the first store item (golf books) was located, transcribed and rated, the target message reappeared on the computer display screen. The computer display screen prompted the subjects to prepare to search for the second item in the target search. The search began from the first store item of the search rather than the main menu. The search began when the information system spoke: "At store item *keyword*, continue searching." Therefore, for multiple searches, subjects had to back-up through menus, or choose to restart from the main menu depending on the proximity of the second store item to the first in the database. Otherwise, the second part of the multiple search proceeded as did the first. Again, after the second store item had been found and the information message heard, the subject pressed the 2 key, transcribed the information message, and completed the rating scales.

Once a subject had finished with the transcription and rating tasks for the first target, a new target was presented on the computer display screen. Subjects were given 15 seconds to read the target and prepare for the next search. The information system then began speaking the main menu options. Subjects proceeded to locate the next target store item(s) and transcribe the information message. The experiment proceeded in this fashion. Each subject completed 16 targets, with single target item subjects searching for 16 store items, and multiple target item subjects searching for 32 store items. Subjects had a mandatory break (minimum 1 miniute) after the first 8 targets. The computer display screen indicated when the subject had completed that part of the experiment.

Upon completion of all target searches, subjects rated the telephone information system on the following measures:

- o ease of use of the information system
- o intelligibility of the computer voice
- o speech rate of the computer voice
- o naturalness of the computer voice
- o system response time to user inputs
- o available time for user inputs
- o complexity of the database organization
- o worth of pause/resume feature
- o worth of repeat speech feature
- o worth of spell-out speech feature
- o worth of wallet guide
- o effect of background music

Examples of the post-experiment ratings are located in see Appendix VI. Each of the measures rated on 7 point bi-polar adjective ratings scales similar to those used during the target searches.

At the conclusion of the experiment, subjects were offered a chance to comment on the system. These comments were recorded in the subjects' diary. Subjects were then debriefed as to the purpose of the experiment, thanked for their participation, and compensated for their time. Table 7 lists the procedures for an experimental session.

Table 7. Experimental Procedures for Screening Study

WELCOME AND ORIENTATION	(~15 mins)
Informed Consent	
Subject Information Questionnaire	
Hearing Test	
INSTRUCTIONS AND PRACTICE	(~20 mins)
Introduction (audio - written)	
Instructions (audio- written)	
Video Instructions	
Telephone Key Instructions (audio - written)	
Subject Recapituates Instructions	
Practice Targets (n=2)	
EXPERIMENTAL TASK	(~30 mins for single target, ~60 mins for multiple targets)
8 Experimental Targets	
Target Search	
Transcription	
Target Ratings	
Break (minimum 1 minute)	
8 Experimental Targets	
Target Search	
Transcription	
Target Ratings	
Post Experimental Ratings	
POST EXPERIMENTAL SESSION	(~15 mins)
Debriefing	
Payment and Dismissal	

Dependent Variables

Twenty-three dependent measures were developed during the engineering analysis phase of the research project. One of the objectives of this screening study was to validate the appropriateness and worth of each dependent measures. Listed below are the 23 dependent measures; 8 objective measures and 15 subjective measures.

Objective Measures
- o target search time ratio
- o target search efficiency ratio
- o invalid keypress average
- o strict message transcription average
- o synonym message transcription average
- o pause/resume selected total*
- o repeat keyword selected total*
- o spell-out keyword selected total*

Subjective Measures
- o message transcription certainty average rating
- o message transcription difficulty average rating
- o difficulty of locating store item average rating
- o ease of use rating
- o computer voice intelligibility rating
- o computer voice naturalness rating
- o computer voice speech rate rating
- o system response time rating
- o user input timeout rating
- o database organization rating
- o pause/resume feature rating*
- o repeat keyword rating*
- o spell-out keyword rating*
- o wallet guide rating*
- o background music rating*

* Only rated when condition is present in experimental session.

Objective Measures. Target search time ratio is an average ratio score of how long it takes the subject to find a target as compared to the minimum time it would take an expert user. The minimum expert target search time is determined by

adding the system time requirements and 0.57 seconds for each menu level selection. System time requirement is the minimum amount of time the system requires to speak the necessary menu items.

The expert selection time of 0.57 seconds is taken from the American Institutes for Research Data Store (Munger, Smith, and Payne, 1962) for an expert user pressing a pushbutton when cued and has a human reliability of 0.999. The time value represents the average amount of time required for an individual to press a single push-button when a stimulus is presented. This scenario can be translated directly to the expert user scenario, because the expert user only needs to push the selection key (single push-button), never makes an error (~0.999 reliability) , and knows the database organization so thoroughly that the expert only selects when the appropriate keyword (stimulus) is spoken by the information system.

As an example, assume the desired target is in the 2x6 database (i.e., 6 levels of menus with 2 items/menu), and the minimum time requirement for the system to speak the necessary menus is calculated to be 22.5 seconds. An expert response time of 3.42 seconds is added to the 22.5 seconds, because the expert makes six selections (1 selection/menu) in the 2x6 database. Therefore, in the 8x2 database, the expert makes two selections and the selection time is 1.14 seconds.

Target search efficiency ratio is an average ratio score of the subject's search efficiency. It is calculated by dividing the minimum number of keypresses necessary to locate the target by the actual number of keypresses used. The purpose of this measure is to assess subject search efficiency.

Invalid keypress average is a measure of incorrect keypresses that are totally inappropriate. Invalid keypresses include pressing the 2 key (information message

key) when still in the menu hierarchy, or selecting a key on the telephone keypad which is not allocated as a command. This measure will provide a measure of how additional command keys affect the number of invalid keypress selections.

Message transcription average is a measure of how well subjects understood and transcribed information messages. There are two scoring metrics for message transcription average; strict scoring and synonym scoring. The test was developed and used by Merva (1987) in an investigation of the effects of speech rate, message repetition and information placement on synthesized speech intelligibility. The procedure is as follows:

> "The beginning and end two words of each transcription were checked for accuracy. Subjects received one point for each correct word in their transcription. Under "strict" scoring, the words in the response needed to be exactly the same as the words in the spoken message to be counted as correct. Under "synonym " scoring, synonyms for the spoken words were accepted as correct (i.e. "luggage" or "baggage" for "package"). Every response was scored under both methods, and the two scoring strategies were analyzed independently" (Merva, 1987, p. 21).

The purpose of this measure is to assess the subject's ability to understand what was said in the information message. Because a subject may understand the spirit of a message without transcribing it verbatim, synonym scoring is also assessed. For example, one subject might transcribe "Holiday specials are available in hardcover and paperback", as "Holiday specials are available in hardback and paperback". In this example, one cannot determine if the subject heard the word "hardcover" incorrectly or simply assimilated the word inaccurately. Regardless, the subject transcribed an answer which was technically correct and did not change the semantics of the message.

Subjective Measures. Message transcription certainty rating is a measure of subjects' certainty of the accuracy of their transcription. After subjects transcribe an information message, subjects will rate how certain they are of their transcription. The certainty rating is on a 7 point bi-polar adjective rating scale, where 1 equals very uncertain and 7 equals very certain. The purpose of this measure is to assess how relative certainty is related to actual transcription accuracy.

Message transcription difficulty rating is a measure of how difficult it was for a subject to understand an information message. After subjects rate their perceived certainty in transcribing an information message, subjects will rate how difficult it was to understand the information message. Again, the difficulty rating is on a 7 point bi-polar adjective rating scale, where 1 equals very difficult and 7 equals very easy. The purpose of this measure is to assess how relative difficulty in understanding messages is related to actual transcription accuracy.

Difficulty of locating the store item rating is a measure of how difficult it was for a subject to search for and locate a store item. After subjects rate their perceived certainty and difficulty in transcribing an information message, subjects will rate how difficult it was to locate the store item. Again, the difficulty rating is on a 7 point bi-polar adjective rating scale, where 1 equals very difficult and 7 equals very easy. The purpose of this measure is to assess how relative difficulty in locating store items in the database.

Ease-of-use rating is recorded at the end of the experiment, after the searching tasks have been completed. The ease-of-use rating is the subject's perception of how easy or hard the telephone information system was to use. This rating is specific to searching throughout the database, and does not include the

ease-of-use for the transcription or rating tasks. Subjects will rate ease-of-use on a 7 point bi-polar adjective scale, where 1 equals very difficult and 7 equals very easy. The purpose of this measure is to assess the effects of all independent variables on subject's perception of how easy the system was to use.

Rating the intelligibility of the computer voice is recorded at the end of the experiment, after the searching task has been completed. The intelligibility rating is the subject's perception of how intelligible the computer voice was for the telephone information system. Subjects will rate intelligibility on a 7 point bi-polar adjective scale, where 1 equals very unintelligible and 7 equals very intelligible. The purpose of this measure is to assess the effects of speech display characteristics (type of voice) on subjects' perception of how easy the computer voice was to understand.

Rating the naturalness of the computer voice is recorded at the end of the experiment, after the searching task has been completed. The naturalness of the computer voice rating is the subject's perception of how natural (or human-like) the computer voice was for the telephone information system. Subjects will rate naturalness on a 7 point bi-polar adjective scale, where 1 equals very unnatural and 7 equals very natural. The purpose of this measure is to assess the effects of speech display characteristics (type of voice) on subjects' perception of how natural the computer voice was as compared to a human voice.

Rating the speech rate of the computer voice is recorded at the end of the experiment, after the searching tasks have been completed. Subjects will rate the speech rate on a 7 point bi-polar adjective scale, where 1 equals very slow and 7 equals very fast. The purpose of this measure is to assess the effects of actual synthesized speech rate on perceived speech rate.

System response time rating is recorded at the end of the experiment, after the searching task has been completed. System response time is how long it takes the system to respond to a subject's command entry (e.g. menu selection, back-up one menu, repeat, etc.). Subjects will rate the system response time on a 7 point bi-polar adjective scale, where 1 equals very slow and 7 equals very fast. The purpose of this measure is to assess the effects of actual system response time on perceived system response time.

Input timeout rating is recorded at the end of the experiment, after the searching task has been completed. Input timeout is the time window between keywords. It is how long the user has to select a keyword or enter a command function. Subjects will rate the system response time on a 7 point bi-polar adjective scale, where 1 equals very little and 7 equals very much. The purpose of this measure is to assess the effects of actual input timeout on perceived input timeout windows.

Database organization rating is recorded at the end of the experiment, after the searching task has been completed. Database organization pertains to the subject's perception of the complexity of the database for telephone information system. Subjects will rate database organization on a 7 point bi-polar adjective scale, where 1 equals very complex and 7 equals very simple. The purpose of this measure is to determine the effects of database organization on subjects' perception of the complexity of the database organization.

Feature ratings (pause/resume, repeat keyword, spell-out keyword, wallet guide, and background music) are requested only when the feature is present in the experimental condition. These ratings are requested at the end of the experiment,

after the searching task has been completed. All of the ratings assess how essential the feature is to the subject. Subjects rate each appropriate feature on 7 point bi-polar scales, where 1 equals not essential and 7 equals absolutely essential. The purpose of these ratings is to compare subjective preferences for system features as they relate to system performance.

RESULTS

A total of 37 subjects were studied; 32 subjects comprise the 32 treatment conditions, 4 other subjects failed the hearing criterion test (their data was retained for possible future studies), and one other subject failed to find a store item within the required 15 minute search time criterion requiring that the subject's treatment condition be replaced.

One of the major challenges in conducting the screening study was determining how best to analyze the data. The decision was made to analyze the data using different analysis techniques. This decision was based on the fact that screening studies do not fit into any traditional convention for analysis. Therefore, by analyzing the data several ways comparisons could be made between analysis techniques to assess what techniques were useful.

Results of the study were analyzed for both the objective and subjective dependent measures. The Statistical Analysis Package (SAS, 1986) was used for the parametric data analysis and was run on the University's IBM 370 mainframe computer. Statview 512+ was used for the nonparametric data analysis and was run on a Macintosh SE.

Dependent measures were recorded using a software metering program prepared especially for this research project. The metering program recorded everything spoken by the telephone information system, as well as every telephone keypad entry by the subject. All events were time-stamped and recorded in chronological order. The package also recorded information message transcriptions and all subjective ratings.

The null hypothesis (H_0) for all analyses was that a variable level did not affect the value of the dependent measure (i.e., H_0: high=low). The alternative hypothesis (H_1) contended that there existed an effect of variable levels on dependent measures but did not suggest a direction (i.e., H_1: high≠low).

Given the objective of the screening study, selecting a decision criterion was not simple. Screening studies do enjoy conventional methods for setting alpha error for hypothesis testing as do traditional full factorial designs. In traditional human factors research, the norm would be to select a very small test criterion (e.g., $p<0.05$, or $p<0.01$). Choosing such a small test criterion guards against Type I errors; rejecting the null hypothesis when in fact the null hypothesis is correct. However for this study, the goal was to identify variables which have no observable effect on user performance. Hence, the test criterion should guard against committing a Type II error; failing to reject the null hypothesis when in fact the null hypothesis is incorrect. Therefore, to reduce the chance of Type II error, the data were analyzed with the decision criterion level set at $p < 0.20$.

For those variables found to have a significant main effect on user performance, post-hoc analyses were performed to identify what differences exist. Because only two levels are being investigated for each variable, post-hoc comparisons were performed by inspection of those effects found to be significant.

Objective Measures

Objective measures were analyzed using several parametric techniques. Traditional multiple analysis of variance (MANOVA) was performed to identify independent variables which were affected significantly across the objective

measures. Subsequently for those variables found to be significant in the MANOVA, analysis of variance (ANOVA) were performed as a deterministic analysis technique, and forward step-wise regression was performed as an inferential analysis technique. To evaluate the uniqueness of the objective measures, a factor analysis (principal components analysis) was performed on the objective measures to determine the correlation between measures across the variables. Frequency counts were also studied to evaluate the utility of certain command features.

Multiple Analysis of Variance. Traditional MANOVA testing was performed for the 16 independent variables across 4 of the objective dependent measures (search time ratio, search efficiency ratio, invalid keypress average, and strict message transcription average). Synonym message transcription average was not included because it was viewed as an alternative metric for scoring transcription average. For each variable, Wilks U-Criterion was calculated and translated into the familiar F-value. The results identified a total of 8 variables as having a significant effect across the 4 dependent measures (see Table 8). These variables were; speech rate, menu organization, number of targets, wallet guide, menu feedback, background music, age of user, and sex of user. It is worth noting that had the test criterion level been set at $p<0.05$, then only menu organization and age of user would have tested significant in the MANOVA.

Table 8. MANOVA Summary Table

Source	df	F*	p
Voice Type	1	0.52	0.7200
Speech Rate	1	2.05	0.1511†
Menu Organization	1	3.73	0.0340†
Number of Targets	1	3.23	0.0512†
System Response Time	1	0.37	0.8249
Input Timeout	1	0.23	0.9137
Pause/Resume	1	0.33	0.8501
Repeat Keyword	1	0.84	0.5272
Spell-out Keyword	1	1.44	0.2800
Wallet Guide	1	2.22	0.1279†
Selection Feedback	1	1.38	0.2995
Menu Feedback	1	2.08	0.1473†
Command Feedback	1	0.66	0.6305
Background Music	1	2.53	0.0954†
Age of Subject	1	3.98	0.0279†
Sex of Subject	1	1.82	0.1891†
Residual Error	15		
Total	31		

*F approximation obtained by conversion using Wilk's Criterion (SAS,1986).

†Test is significant at $p<.20$

Analysis of Variance. Subsequently, ANOVAs were performed for those variables identified in the MANOVA analysis. The results showed that each of the variables identified in the MANOVA was also significant in at least one ANOVA (see Table 9). In the search time ratio ANOVA, the variables age, and wallet guide were significant (see Table 10). Subjects with a wallet guide had an average search time ratio that was higher (i.e., shorter search time) than there counterparts without a wallet guide. For subject age, college age students had an average search time ratio that was higher than the middle age subjects. These average scores are summarized in Table 11.

In the search efficiency ratio ANOVA, the results were similar to the first ANOVA, with the variables menu organization, age, and wallet guide all being significant (see Table 12). For menu organization, subjects with the 8x2 database had an average search efficiency ratio that was higher (i.e., more efficient) than those subjects with the 2x6 database. Subjects with a wallet guide or who were college age had higher search efficiency ratio scores than their respective counterparts. These average scores are presented in Table 11.

In the invalid keypress average ANOVA, menu organization, number of targets, age, menu feedback, and music were significant at a test criterion of $p<.20$ (see Table 13). For menu organization, subjects with the 8x2 database had an invalid keypress average higher (i.e., more invalid keypresses) than those subjects with the 2x6 database. Subjects that had multiple searches had a higher invalid keypress average (per store item search) than those subjects with single target searches. For subject age, middle age subjects had a higher invalid keypress average than the college age subjects. Subjects that had menu feedback had a

Table 9. Summary of Results for MANOVA/ANOVA Tests

Variable	MANOVA	STR	SER	IKA	SMTA	SYMTA
Speech Rate	*				*	*
Menu Organization	**		*	**	*	*
No. Targets	*			*		
Wallet Guide	*	*	*		*	*
Menu Feedback	*				*	*
Background Music	*			*		
Subject Age	**	*	**	*	**	**
Subject Sex	*				*	*

* significant at p<.20
** significant at p<.05

ANOVA Tests

STR = Search Time Ratio

SER = Search Efficiency Ratio

IKA = Invalid Keypress Average

SMTA = Strict Message Transcription Average

SYMTA = Synonym Message Transcription Average

Table 10. Search Time Ratio ANOVA Summary Table

Source	df	F	p
Menu Organization	1	0.78	0.3922
Number of Targets	1	0.12	0.7392
Sex of Subject	1	0.02	0.8821
Age of Subject	1	4.30	0.0557†
Wallet Guide	1	8.62	0.0102†
Background Music	1	0.29	0.5960
Speech Rate	1	1.16	0.2978
Voice Type	1	0.00	0.9846
System Response Time	1	0.62	0.4444
Input Timeout	1	0.22	0.6447
Pause/Resume	1	0.52	0.4818
Spell-out Keyword	1	0.42	0.5261
Repeat Keyword	1	0.03	0.8644
Selection Feedback	1	4.66	0.0475
Menu Feedback	1	1.54	0.2343
Command Feedback	1	0.71	0.4117
Residual Error	15		
Total	31		

†Test is significant at $p<.20$

Table 11. Summary of Averages for ANOVA Tests

Variable	STR	SER	IKA	SMTA	SYMTA
Speech Rate					
180				3.43	3.54
240				3.10	3.25
Menu Organization					
8x2		.899	.156	3.36	3.49
2x6		.845	.041	3.16	3.29
No. Targets					
1			.047		
2			.150		
Wallet Guide					
Available	.906	.918		3.36	3.49
Not Available	.756	.826		3.16	3.29
Menu Feedback					
Available			.131	3.13	3.28
Not Available			.066	3.40	3.51
Background Music					
Not Available			.053		
Available			.145		
Subject Age					
College	.884	.919	.059	3.41	3.51
Middle	.779	.825	.139	3.12	3.28
Subject Sex					
Male				3.13	3.27
Female				3.39	3.51

ANOVA Tests
STR = Search Time Ratio (larger numbers are faster times)
SER = Search Efficiency Ratio (larger numbers are more efficient searches)
IKA = Invalid Keypress Average (larger numbers are more invalid keypresses)
SMTA = Strict Message Transcription Average (larger numbers are better trans.)
SYMTA = Synonym Message Transcription Average (larger numbers are better transcription)

Table 12. Search Efficiency Ratio ANOVA Summary Table

Source	df	F	p
Menu Organization	1	2.00	0.1781†
Number of Targets	1	0.27	0.6090
Sex of Subject	1	0.24	0.6283
Age of Subject	1	6.15	0.0255†
Wallet Guide	1	5.72	0.0303†
Background Music	1	1.22	0.2877
Speech Rate	1	0.80	0.3844
Voice Type	1	0.06	0.8124
System Response Time	1	0.49	0.4964
Input Timeout	1	0.34	0.5666
Pause/Resume	1	0.37	0.5507
Spell-out Keyword	1	0.15	0.6995
Repeat Keyword	1	0.05	0.8331
Selection Feedback	1	5.83	0.0290
Menu Feedback	1	1.34	0.2645
Command Feedback	1	1.33	0.2662
Residual Error	15		
Total	31		

†Test is significant at $p<.20$

Table 13. Invalid Keypress Average ANOVA Summary Table

Source	df	F	p
Menu Organization	1	7.17	0.0172†
Number of Targets	1	5.78	0.0296†
Sex of Subject	1	0.59	0.4525
Age of Subject	1	3.46	0.0826†
Wallet Guide	1	0.02	0.8934
Background Music	1	4.55	0.0499†
Speech Rate	1	0.74	0.4022
Voice Type	1	1.50	0.2395
System Response Time	1	0.46	0.5065
Input Timeout	1	0.25	0.6250
Pause/Resume	1	0.05	0.8236
Spell-out Keyword	1	2.82	0.1139
Repeat Keyword	1	0.10	0.7552
Selection Feedback	1	0.91	0.3558
Menu Feedback	1	2.24	0.1551†
Command Feedback	1	0.35	0.5641
Residual Error	15		
Total	31		

†Test is significant at $p<.20$

higher invalid keypress average than those that did not have menu feedback. Also, subjects that listened to background music had a higher invalid keypress average than those subjects that did not hear music. These average scores are summarized in Table 11.

In the strict message transcription average ANOVA, menu organization, sex, age, wallet guide, speech rate, and menu feedback were significant (see Table 14). Subjects with the 8x2 database had a higher strict message transcription average than those with the 2x6 database. College age subjects also had a higher transcription average than middle age subjects. For sex of subject, females had higher transcription average scores than males. Those subjects that heard DECtalk speak at 180 wpm had a higher strict message transcription score than those subjects that heard the speech synthesizer at 240 wpm. For menu feedback, those subjects that did not receive menu feedback had a higher strict message transcription than those subjects that did have menu feedback. These average scores are summarized in Table 11.

An ANOVA was calculated for synonym message transcription average as an alternative metric for message transcription. The results of the synonym message transcription average ANOVA were identical to strict message transcription average ANOVA, with menu organization, sex, age, wallet guide, speech rate, and menu feedback again being significant (see Table 15). Again, subjects who had the 8x2 database, were college age, were female, heard the synthesizer at 180 wpm, or did not receive menu feedback had higher synonym message transcription scores than their respective counterparts. The average scores are summarized in Table 11.

Table 14. Strict Message Transcription Average ANOVA Summary

Source	df	F	p
Menu Organization	1	2.86	0.1116†
Number of Targets	1	0.02	0.9002
Sex of Subject	1	4.57	0.0494†
Age of Subject	1	5.57	0.0322†
Wallet Guide	1	2.75	0.1179†
Background Music	1	0.80	0.3858
Speech Rate	1	7.01	0.0183†
Voice Type	1	1.71	0.2106
System Response Time	1	0.00	0.9750
Input Timeout	1	0.20	0.6615
Pause/Resume	1	1.18	0.2952
Spell-out Keyword	1	2.97	0.1055
Repeat Keyword	1	2.86	0.1116
Selection Feedback	1	0.41	0.5331
Menu Feedback	1	4.84	0.0438†
Command Feedback	1	0.03	0.8754
Residual Error	15		
Total	31		

†Test is significant at $p<.20$

Table 15. Synonym Message Transcription Average ANOVA Summary

Source	df	F	p
Menu Organization	1	3.69	0.0738†
Number of Targets	1	0.49	0.4933
Sex of Subject	1	5.08	0.0396†
Age of Subject	1	4.59	0.0489†
Wallet Guide	1	3.69	0.0738†
Background Music	1	0.60	0.4498
Speech Rate	1	7.48	0.0153†
Voice Type	1	1.15	0.3008
System Response Time	1	0.07	0.7994
Input Timeout	1	0.20	0.6638
Pause/Resume	1	0.55	0.4712
Spell-out Keyword	1	1.67	0.2154
Repeat Keyword	1	4.44	0.0524
Selection Feedback	1	0.49	0.4933
Menu Feedback	1	4.59	0.0489†
Command Feedback	1	0.01	0.9421
Residual Error	15		
Total	31		

†Test is significant at $p<.20$

Regression Analysis. Forward, step-wise regression analysis was also performed for those variables identified in the MANOVA analyses. The purpose of this analysis was to determine if additional information regarding variable effects could be gained from an inferential (i.e., predictive) analysis technique. First order regression models were constructed which only considered main effects of the 16 independent variables; other sources of variance were pooled in the residual term.

The results of the regression analysis echoed the results of the ANOVA analyses (see Tables 16 through 20). For the five tests, R^2 values ranged from 0.45 to 0.63 with only main effects considered in the model (see Table 21). Search time ratio results were closely related to search efficiency ratio results, and strict message transcription average results were identical to synonym message transcription average results.

For each forward, step-wise regression model, the significant variables are identical to the results of the respective ANOVA test. For search time ratio, two variables, age and wallet guide, were significant in the regression model (see Table 16). In the search efficiency ratio regression model, menu organization, age, and wallet guide were significant (see Table 17). In the invalid keypress average regression model, menu organization, number of targets, age, and music were significant (see Table 18). In the strict message transcription average regression model, menu organization, sex, age, wallet guide, speech rate, and menu feedback were significant (see Table 19). Finally, the regression model for the synonym message transcription average was, again, identical to strict message transcription average, with menu organization, sex, age, wallet guide, speech rate, and menu feedback all being significant (see Table 20).

Table 16. Stepwise Regression Summary Table for Search Time Ratio

Source	df	F	p
Regression	3	7.66	0.0070
Wallet Guide	1	8.57	0.0066
Selection Feedback	1	5.25	0.0294
Subject Age	1	5.62	0.0249
Error	28		

Regressors	B Value
Intercept	0.9298
Wallet Guide	-0.1055
Selection Feedback	0.1493
Subject Age	-0.1098

$R^2 = 0.45$

Table 17. Stepwise Regression Summary Table for Search Efficiency Ratio

Source	df	F	p
Regression	4	6.14	0.0012
Menu Organization	1	2.49	0.1264
Wallet Guide	1	6.77	0.0147
Selection Feedback	1	5.76	0.0231
Subject Age	1	5.24	0.0293
Error	27		

Regressors	B Value
Intercept	0.9700
Menu Organization	0.0090
Wallet Guide	0.0919
Selection Feedback	-0.0928
Subject Age	-0.0953

$R^2 = 0.48$

Table 18. Stepwise Regression Summary Table for Invalid Keypress Average

Source	df	F	p
Regression	6	5.43	0.0010.
Menu Organization	1	5.54	0.0254
Number of Targets	1	5.07	0.0320
Spell-out Keyword	1	3.30	0.0809
Menu Feedback	1	2.81	0.1064
Background Music	1	4.47	0.0436
Subject Age	1	3.73	0.0640
Error	25		

Regressors	B Value
Intercept	-0.3988
Menu Organization	0.0192
Number of Targets	0.1035
Spell-out Keyword	-0.0722
Menu Feedback	0.0644
Background Music	0.0918
Subject Age	0.0801

$R^2 = 0.57$

Table 19. Stepwise Regression Summary Table for Strict Message Transcription Average

Source	df	F	p
Regression	8	4.97	0.0012
Speech Rate	1	4.60	0.0403
Subject Age	1	4.02	0.0543
Menu Feedback	1	3.84	0.0601
Subject Sex	1	4.01	0.0554
Spell-out Keyword	1	2.77	0.1078
Menu Organization	1	2.86	0.1030
Repeat Keyword	1	3.11	0.0907
Wallet Guide	1	3.27	00835.
Error	23		

Regressors	B Value
Intercept	4.3704
Menu Organization	0.0345
Subject Sex	0.2617
Subject Age	-0.2891
Wallet Guide	0.2031
Speech Rate	-0.0054
Spell-out Keyword	-0.2109
Repeat Keyword	0.2070
Menu Feedback	0.2695

$R^2 = 0.63$

Table 20. Stepwise Regression Summary Table for Synonym Message Transcription Average

Source	df	F	p
Regression	7	5.69	0.0006
Speech Rate	1	4.84	0.0356
Subject Sex	1	3.57	0.0687
Subject Age	1	3.51	0.0714
Menu Feedback	1	3.87	0.0595
Repeat Keyword	1	4.18	0.0512
Menu Organization	1	3.86	0.0607
Wallet Guide	1	4.38	0.0471
Error	24		

Regressors	B Value
Intercept	3.9186
Menu Organization	0.0339
Subject Sex	0.2383
Subject Age	-0.2266
Wallet Guide	0.2031
Speech Rate	-0.0048
Repeat Keyword	0.2226
Menu Feedback	-0.0066

$R^2 = 0.62$

Table 21. R^2 values for Forward, Stepwise Regression Models

Regression Model	R^2 Value
Search Time Ratio	0.45
Search Efficiency Ratio	0.48
Invalid Keypress Average	0.57
Strict Transcription Average	0.63
Synonym Transcription Average	0.62

Principal Components Analysis. A principal components analysis (factor analysis) was performed on the 5 objective measures discussed above. This analysis was performed to identify principal components for these dependent measures that represented a significant proportion of the variance, calculate the weighting of each measure in a significant component, and determine the correlation between these dependent measures. The Kaiser (1958) criterion of retaining only those components with an eigenvalue greater than or equal to 1.0 was used in the analysis. Table 22 lists the eigenvalues and proportion of variance accounted for by each component. The analysis identified two combined measures. By inspection, the first measure can be interpreted as an efficiency measure because it includes searching and transcription scores. The other combined measure can be termed as an accuracy measure because it includes keying and transcription measures.(see Table 23). The efficiency component is a measure of how proficient the subject is at using the information system. The accuracy component is a measure of how accurate the subject is at keying entries and transcribing messages.

Ratio scores were highly correlated, and not surprisingly, message transcription average scores were also highly correlated. The search time and efficiency ratio measures had a positive effect on the efficiency score and a negative effect on the accuracy score. The invalid keypress average was not correlated with any other measure, and had a weak negative effect on the efficiency measure and a strong positive effect on the accuracy measure. The message transcription measures had a negative effect on the efficiency score and a positive effect on the accuracy score. These findings suggest strong relationship between the ratio scores and between the transcription averages.

Table 22- Eigenvalues and Proportion of Variance for Each Component

Component	Eigenvalue	Proportion of Variance
A (Efficiency)	2.91*	0.5827
B (Accuracy)	1.34*	0.2689
C	0.68	0.1353
D	0.05	0.0097
E	0.02	0.0033

* Eigenvalue > 1.0

Table 23. Principal Components Analysis Summary Table

CORRELATIONS

	STR	SER	IKA	STA	SYTA
STR	--				
SER	0.951	--			
IKA	-3.20	-2.69	--		
SMTA	0.474	0.441	0.032	--	
SYMTA	0.481	0.429	0.046	0.980	--

COMPONENT WEIGHTING

	EFFICIENCY	ACCURACY
STR	0.874	-0.375
SER	0.845	-0.378
IKA	-0.231	0.756
SMTA	0.833	0.490
SYMTA	0.830	0.499

Frequency Count Analysis. Frequency of use counts were calculated for each command feature available to subjects. Specifically, frequency counts were tabulated for pause/resume, repeat keyword, and spell-out keyword. The purpose of counting was to provide a simple method for assessing the utility of each command feature based on the frequency of use throughout the experimental session. The data were also helpful for supporting or contradicting any analytical test results regarding these three variables.

Frequency counts for the three variables are summarized in Table 24. Remember, each feature was only available to half (16) of the subjects. As shown in the table, pause/resume was never used by any subject. Repeat keyword was used by two subjects with one subject using the feature nine times and the other subject using the feature two times. For spell-out keyword, this feature was used by two subjects with each subject using the feature one time.

Table 24. Frequency Counts for Telephone Keypad Command Features

Command Feature	Number of Subjects	Frequency of Use/Subject
Pause/Resume	0	0/0
Repeat Keyword	2	9/1 2/2
Spell-out Keyword	2	1/1 1/2

Subjective Measures

Subjective measures were analyzed using the nonparametric Mann-Whitney U Test to determine the effects of the independent variables on different subjective rating scores. In addition, frequency distributions for treatment-specific ratings (e.g., wallet guide rating) were analyzed to determine user preferences.

Mann-Whitney U Tests. The Mann-Whitney U nonparametric test was selected because it has the strongest power of any nonparametric test for ordinal data with k-variables (Siegel, 1956). For each subject, median scores were calculated for the 10 subjective ratings listed below:

- message transcription certainty rating
- message transcription difficulty rating
- store item search difficulty rating
- ease of use rating
- computer voice intelligibility rating
- computer voice naturalness rating
- computer voice speech rate rating
- menu organization rating
- system response time rating
- user input timeout rating

A total of 55 Mann-Whitney U Tests were performed with 14 of the tests finding a significant effect for an independent variable. The decision criterion for all tests was, as in parametric testing, set at $p<0.20$, meaning that a variable was significant if it had a calculated $U<95$ (Siegel, 1956). Appendix VII contains a listing of all 44 Mann-Whitney Tests.

In the first set of tests, Mann-Whitney tests were performed on two general ratings for each of the 16 variables. These ratings were:

- store item search difficulty rating
- ease-of-use rating

All 16 variables were tested separately because responses to these ratings could likely have been affected by any one of the 16 variables. Table 25 summarizes the findings of these Mann-Whitney U Tests.

For store item search difficulty rating, background music, menu feedback, and command feedback variables were significant ($p<.20$). Frequency distributions for store item search difficulty for these three variables are presented in Figures 1 through 3. Subjects that had menu or command feedback rated search difficulty as being more difficult than subjects without menu or command feedback. Subjects that listened to background music rated search difficuluty as being more difficult than those subjects that did not listen to background music.

For ease-of-use rating, five variables had a significant effect on rating scores. Menu organization, number of targets, pause/resume, menu feedback, and subject age all had a significant effect. Frequency distributions for ease-of-use rating for the five variables are depicted in Figures 4 through 8. Subjects that used the 8x2 database rated the system easier to use than those with the 2x6 database. Suprisingly, subjects that searched for two store items rated the system easier to use than those subjects that searched for one store item per target. The effects of Pause/Resume on the ease of use rating are not clear. Subjects that did not have menu feedback rated the system easier to use than theri counter parts. Finally, college age subjects found the system easier to use than middle age subjects.

Table 25. Summary of Significant Results from Mann-Whitney U Tests

RATING (Variable)	U-Value*
GENERAL RATINGS	
(Variables: All 16 Variables)	
Store Item Search Difficulty Rating	
Menu Feedback	94
Command Feedback	78
Background Music	94
Ease-of-Use Rating	
Menu Organization	87.5
Number of Targets	86
Pause/Resume	93
Menu Feedback	88
Age	80
VOICE CHARACTERISTIC RATINGS	
(Variables: Voice Type, Speech Rate, Subject Age, and Subject Sex)	
Message Transcription Certainty Rating	
Voice Type	85
Message Transcription Difficulty Rating	
Voice Type	91.5
Computer Voice Intelligibility Rating	
Subject Age	55
Computer Voice Naturalness Rating	
Voice Type	90.5
Subject Age	81.5
Computer Voice Speech Rate Rating	
Speech Rate	65
Subject Age	75.5
Subject Sex	73.5

Table 25. Summary of Significant Results from Mann-Whitney U Tests - Continued

RATING (Variable)	U-Value*
VARIABLE SPECIFIC RATINGS	
(Variables: Menu Organization, System Response Time, User Input Timeout)	
Menu Organization Rating	
Menu Organization	64.5
System Response Time Rating	
-No significant effect	
User Input Timeout Rating	
-No significant effect	

* Test significant at $p<.20$ when $U<95$

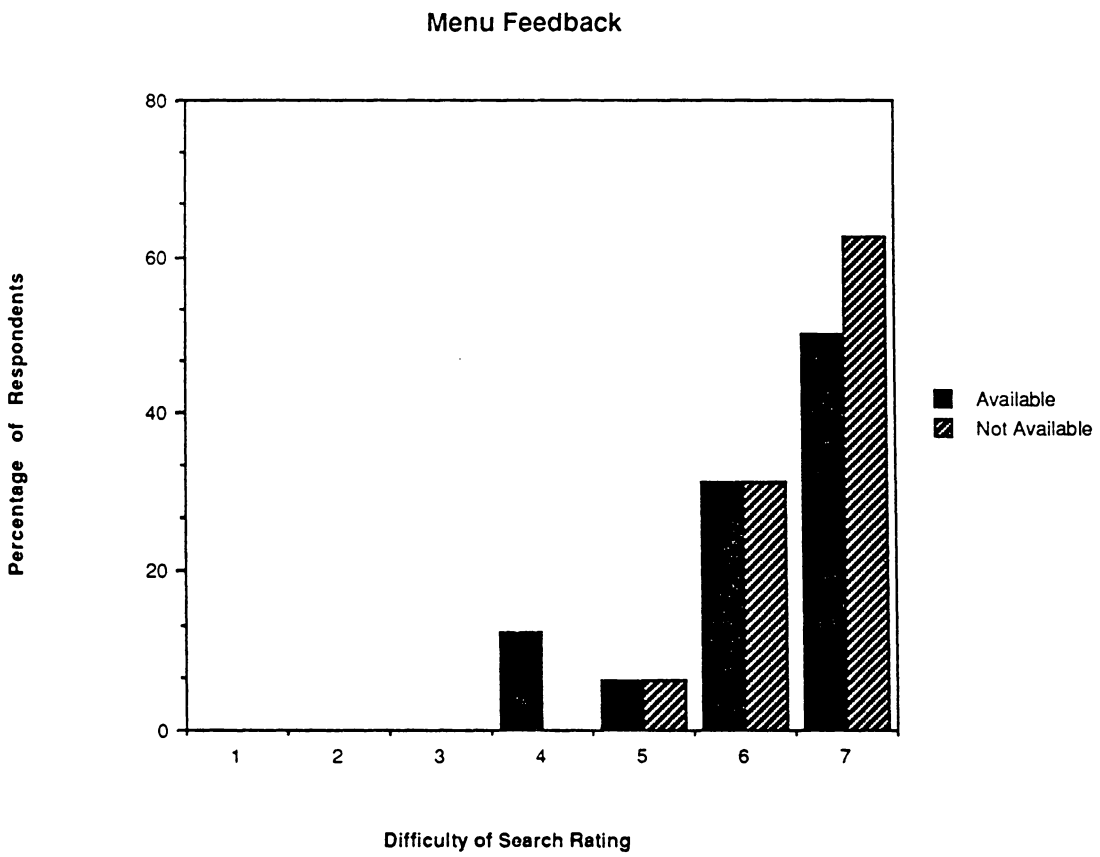


Figure 1 - Frequency Distributions for Store Item Search Difficulty Rating by Menu Feedback

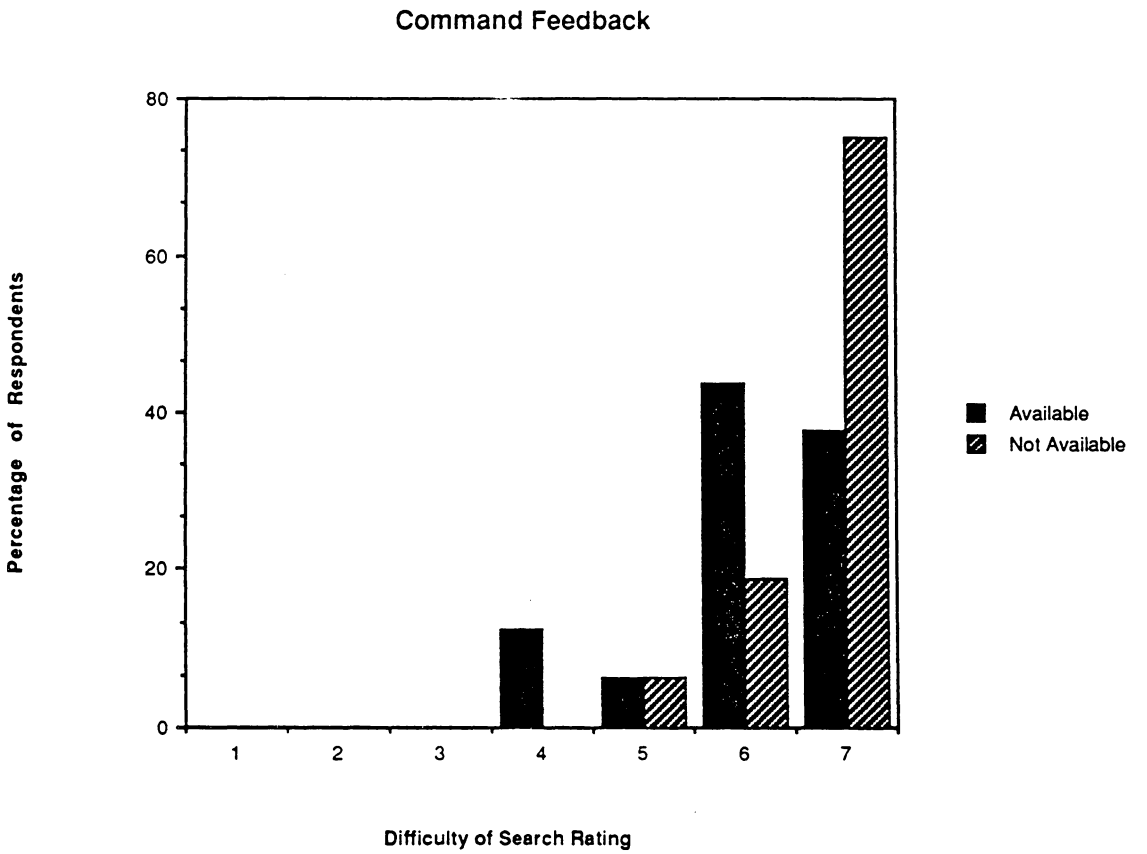


Figure 2 - Frequency Distributions for Store Item Search Difficulty Rating by Command Feedback

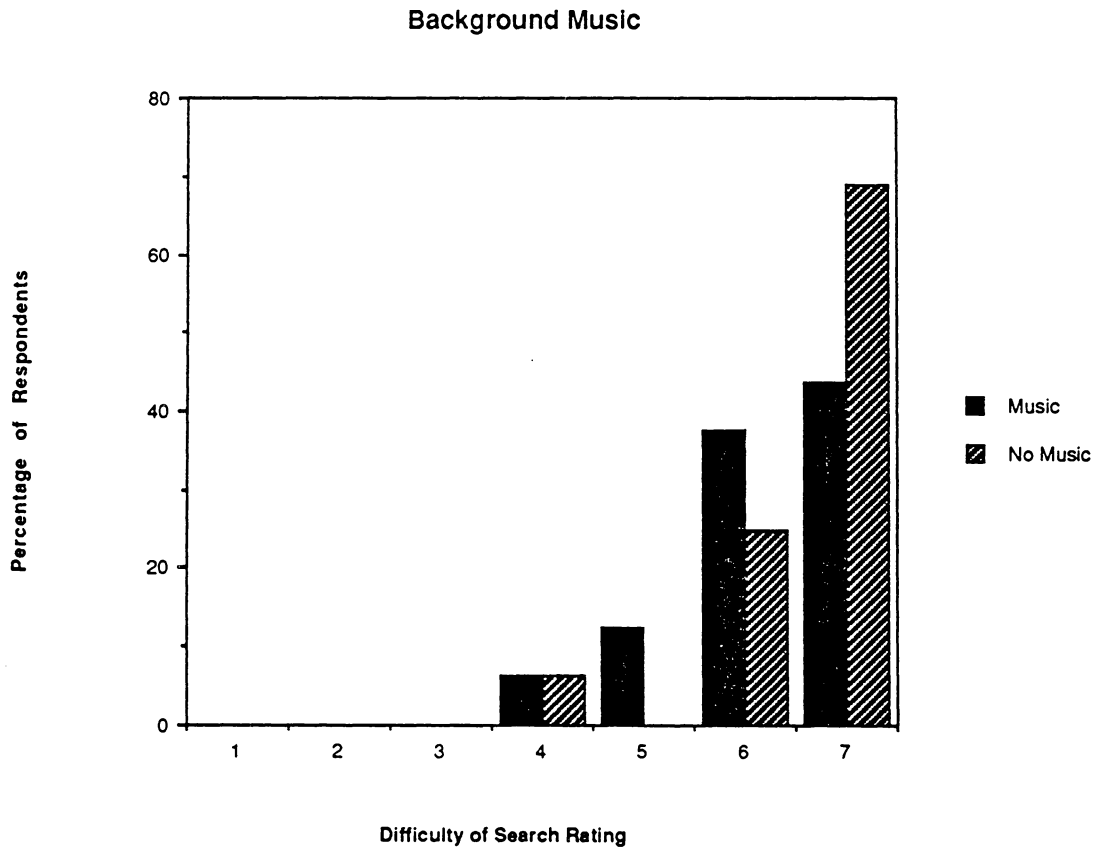


Figure 3 - Frequency Distributions for Store Item Search Difficulty Rating by Background Music

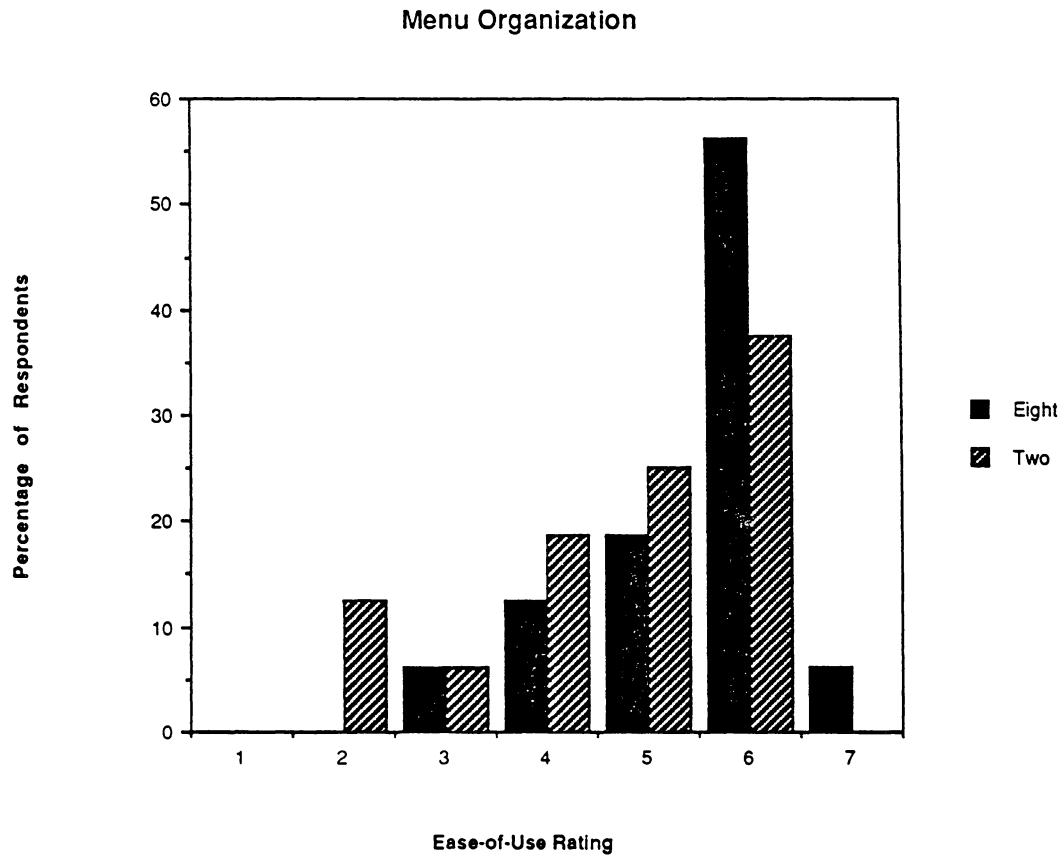


Figure 4 - Frequency Distributions for Ease-of-Use Rating by Menu Organization

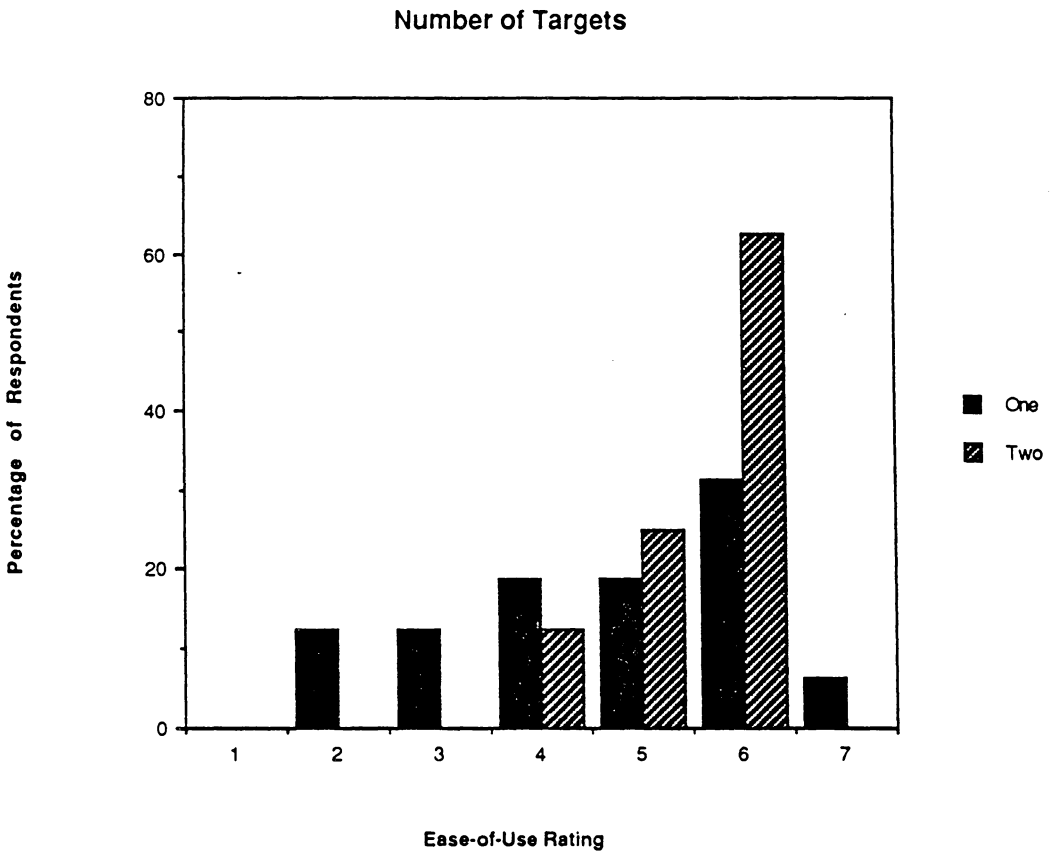


Figure 5 - Frequency Distributions for Ease-of-Use Rating by Number of Targets

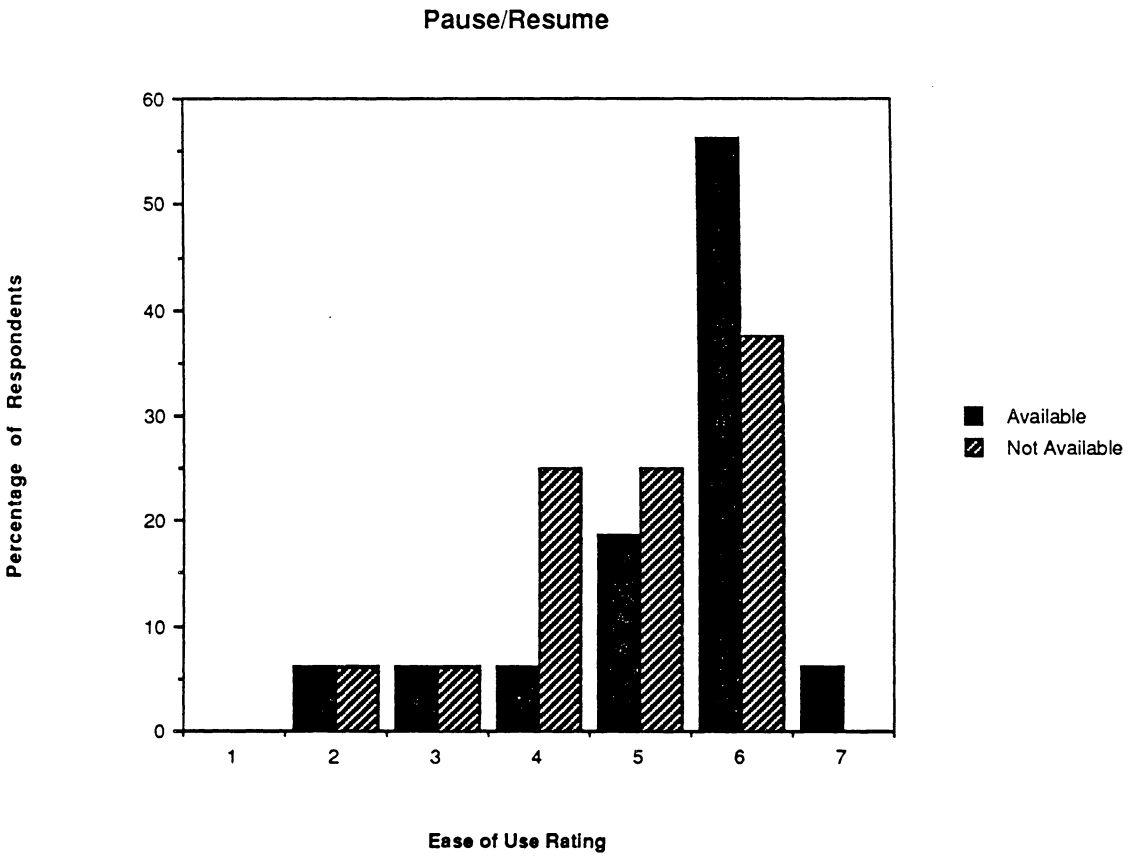


Figure 6 - Frequency Distributions for Ease-of-Use Rating by Pause/Resume

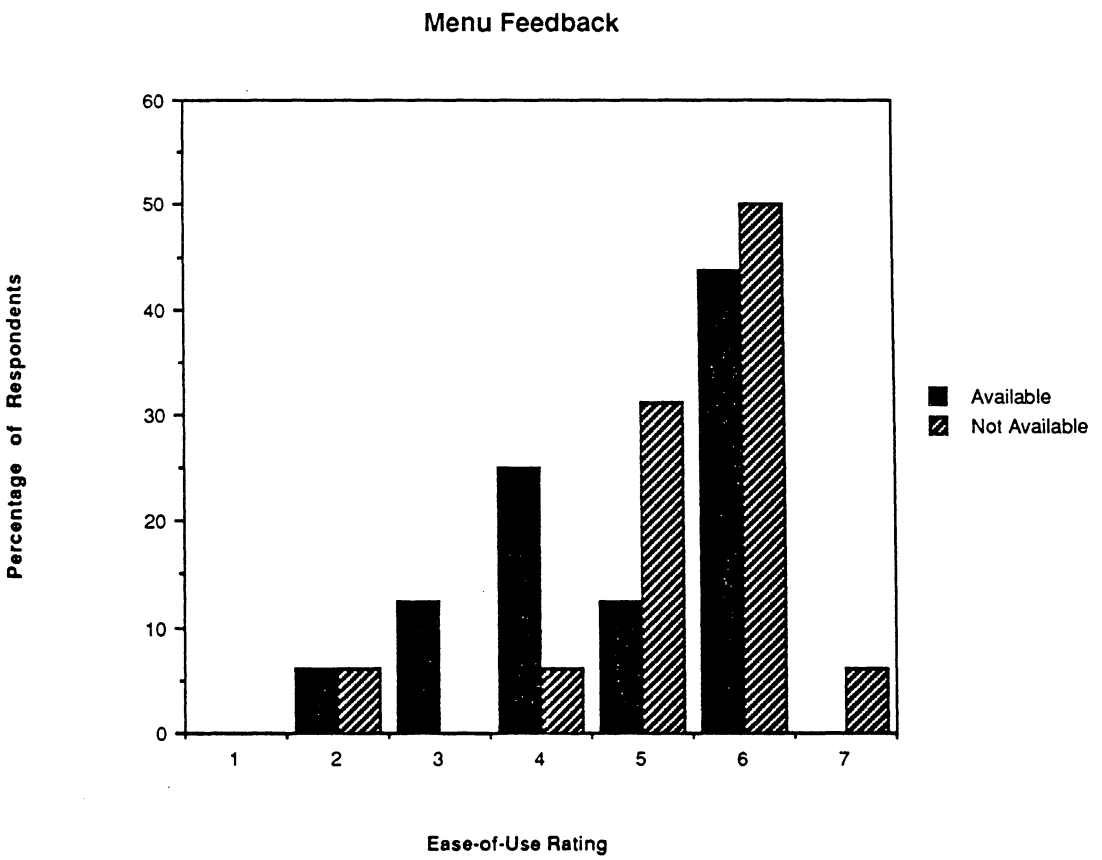


Figure 7 - Frequency Distributions for Ease-of-Use Rating by Menu Feedback

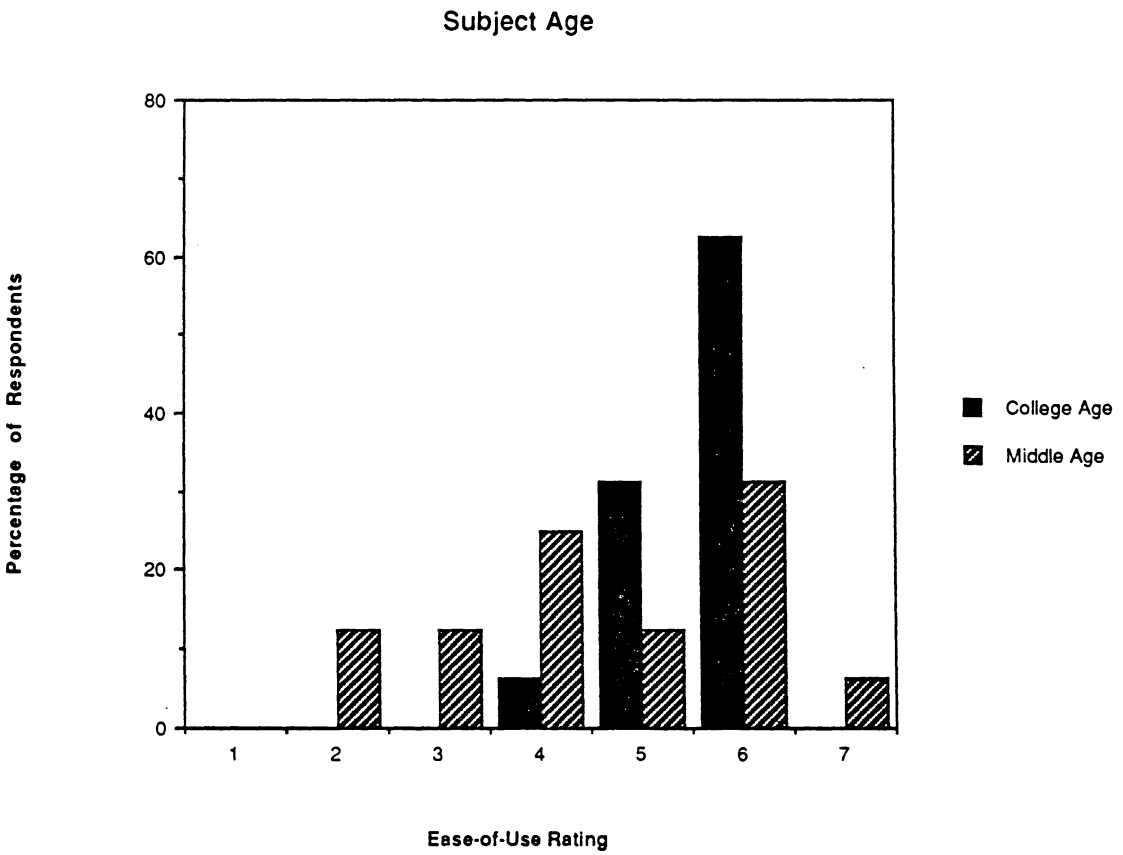


Figure 8 - Frequency Distributions for Ease-of-Use Rating by Subject Age

In the second set of Mann-Whitney U Tests subjective ratings pertaining to synthetic speech voice characteristics were tested for the independent variables voice type, speech rate, subject age, and subject sex. The five voice characteristic ratings tested for each of the four variables were:

- message transcription certainty rating
- message transcription difficulty rating
- computer voice intelligibility rating
- computer voice naturalness rating
- computer voice speech rate rating

The four independent variables directly influence subjects' ratings on voice characteristic measures. Voice type and speech rate are self-evident; each variable level produces a unique voice characteristic. Subject age and sex are demographic variables which can be used for categorizing the opinions (or ratings in this case) for groups of people. Voice characteristic ratings are driven by subjects' perception of a machine generated voice as compared to the familiar sounds of human voices. Thus, age and sex are two variables which could affect subjects' perception of what they heard and what they think about the voice they heard. The remaining twelve variables were not tested for these measures because they did not have a discernable relationship to voice characteristics, and therefore, should not be tested.

The results of the voice characteristic Mann-Whitney U-Tests are summarized in Table 25. For message transcription certainty and message transcription difficulty ratings, voice type was the only variable to have a significant

effect on these ratings. Figures 9 and 10 present frequency distributions for the ratings by voice type. Subject age was the only variable to have a significant effect on computer voice intelligibility, while voice type and subject age had an effect on computer voice naturalness (see Figures 11 and 12 for frequency distributions). Also, speech rate, subject age, and subject sex all had a significant effect on the speech rate rating . Figures 13, 14, 15 present the frequency distributions of speech rate, subject age, and subject sex for the speech rate rating.

Finally, Mann-Whitney U Tests were run for three specific and unique ratings with each addressing a specific variable. These three ratings were:

- menu organization rating
- system response time rating
- user input rating

Each rating was tested for the specific variable associated with the rating. Respectively, these are menu organization, system response time, and input timeout. Again, the results are summarized in Table 25. Menu organization was the only variable which had a significant effect on its corresponding rating. The other two tests found that the corresponding variable did not have a significant effect on the rating. Figure 16 presents the frequency distribution for menu organization by menu organization levels. Subjects with the 2x6 database rated the menu organization to be more complex than those subjects that had the 8x2 database.

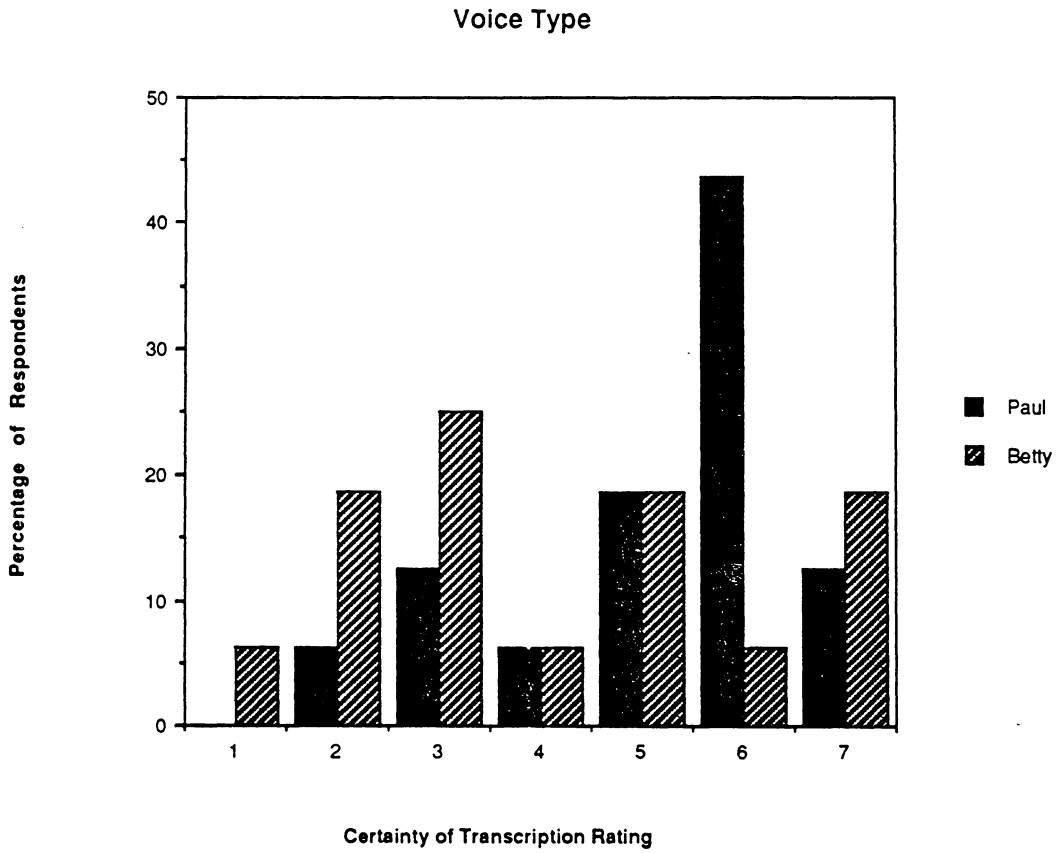


Figure 9 - Frequency Distributions for Message Transcription Certainty Rating by Voice Type

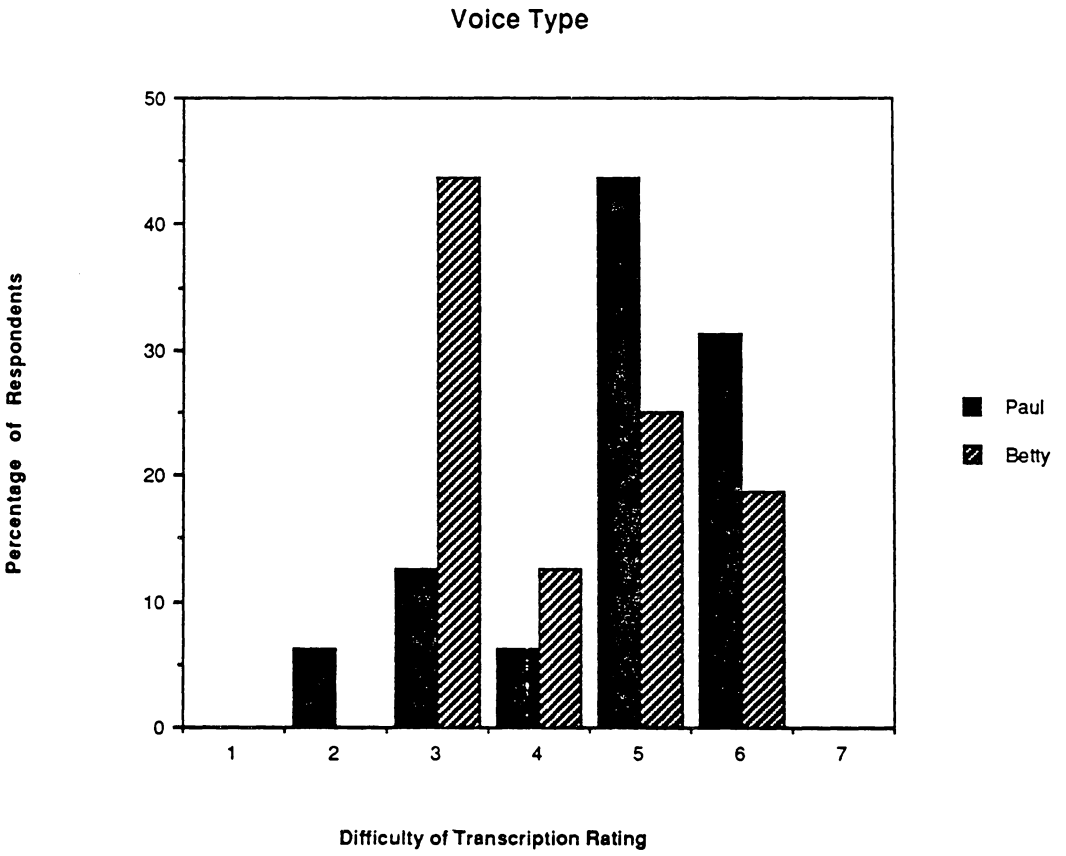


Figure 10 - Frequency Distributions for Message Transcription Difficulty Rating by Voice Type

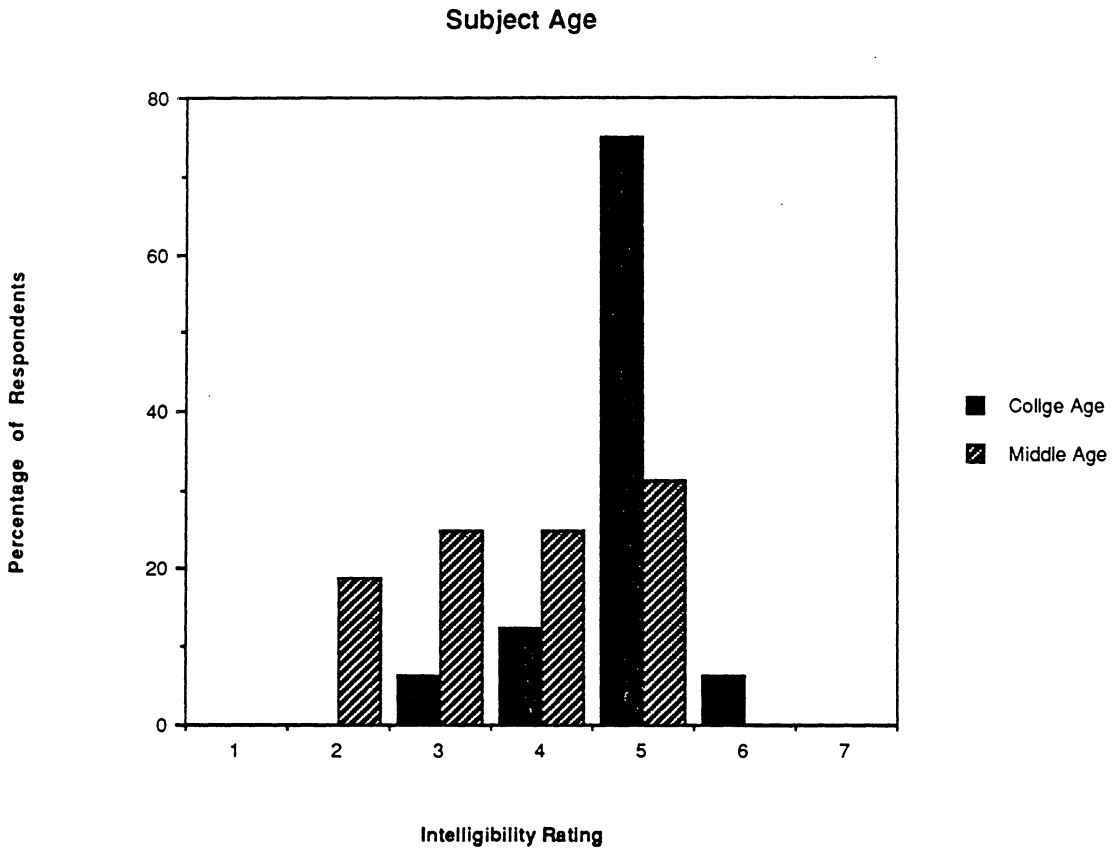


Figure 11 - Frequency Distributions for Computer Voice Intelligibility Rating by Subject Age

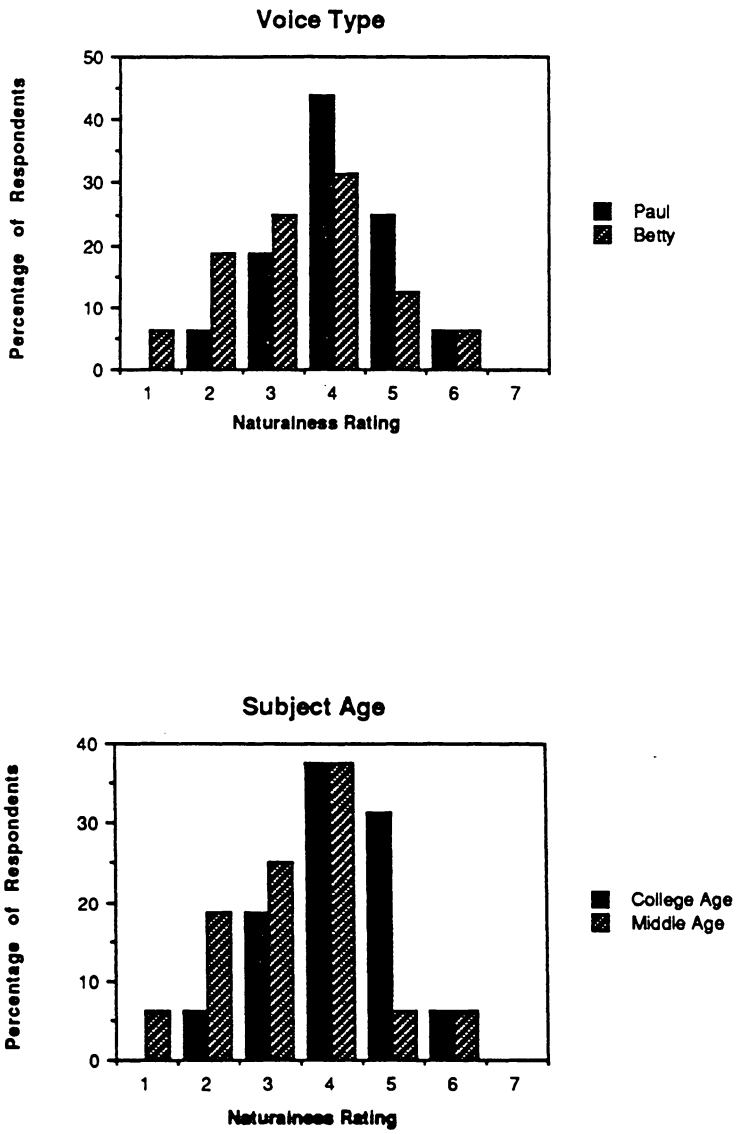


Figure 12 - Frequency Distributions for Computer Voice Naturalness Rating by Voice Type and Subject Age

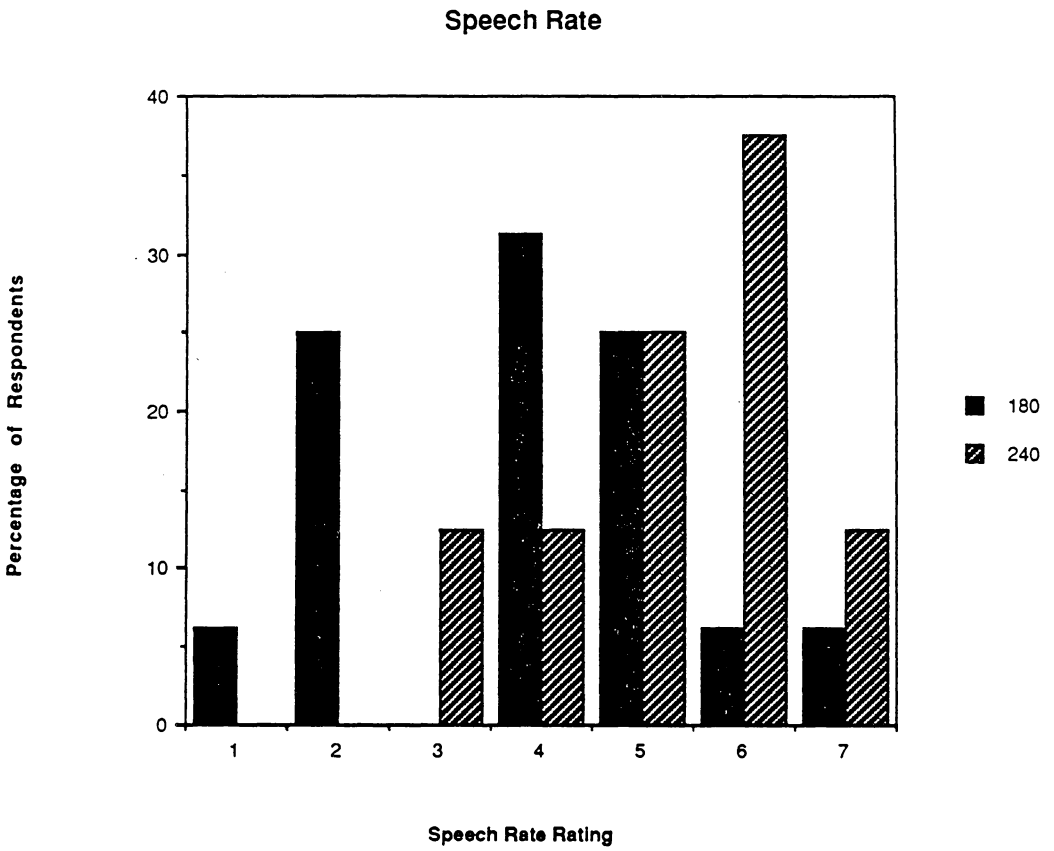


Figure 13 - Frequency Distributions for Computer Voice Speech Rate Rating by Speech Rate

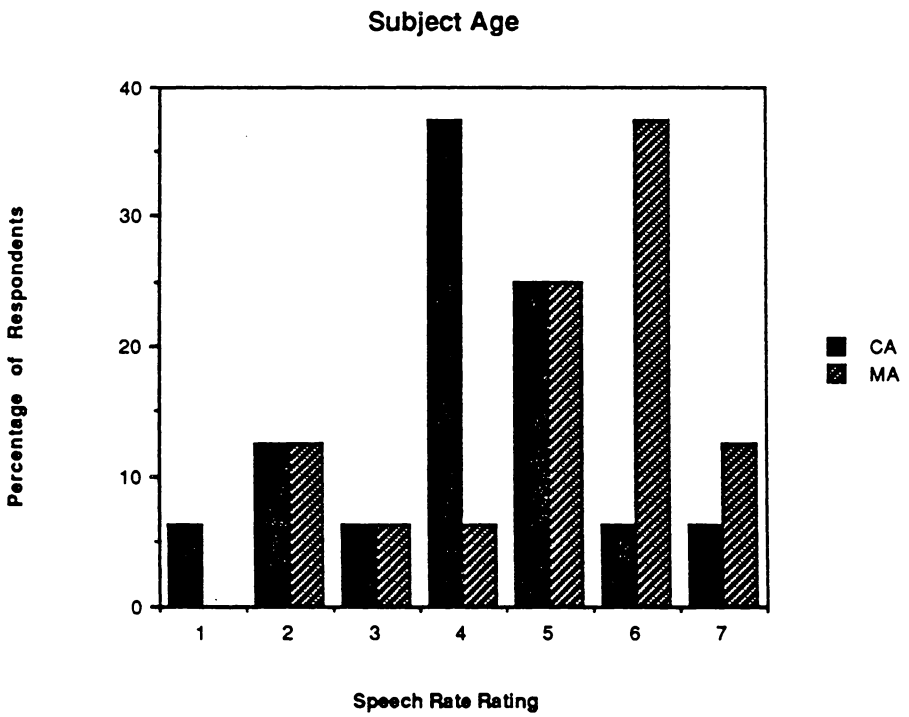


Figure 14 - Frequency Distributions for Computer Voice Speech Rate Rating by Subject Age

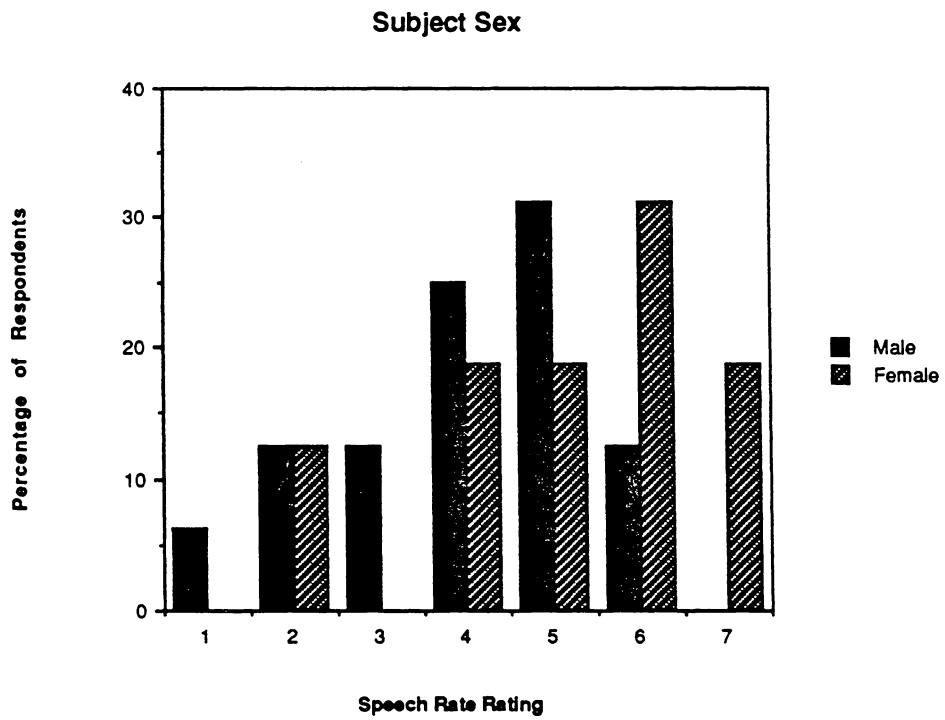


Figure 15 - Frequency Distributions for Computer Voice Speech Rate Rating by Subject Sex

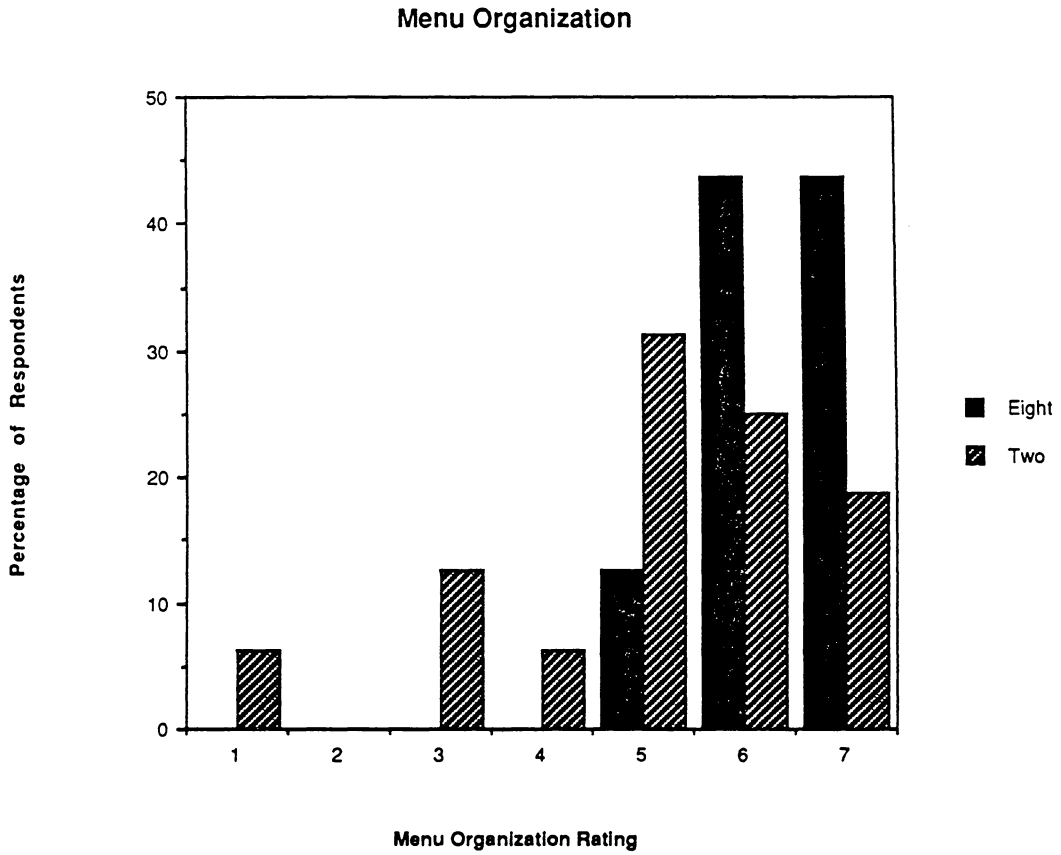


Figure 16 - Frequency Distributions for Menu Organization Rating by Menu Organization

Frequency Distributions for treatment-specific ratings. Several of the variables had a system feature that was either available or not available during the experimental session(e.g. wallet guide). The variable ratings which were treatment specific were:

- pause/resume rating
- repeat keyword rating
- spell-out keyword rating
- wallet guide rating
- background music

When one of these variables was available to a subject, the subject rated the essentiality of the variable as a system feature. Subjects that did not have the feature were not presented with the rating. Thus, only half of the subjects rated the variable. This, of course, eliminated any statistical testing of the effects of variable levels on the respective rating. However, inspection of frequency distribution provides useful information regarding subject preferences.

In Figures 17 through 21, frequency distributions for the 5 ratings are presented. For those subjects that rated pause/resume, 75% of the respondents found the feature very unessential. This corresponds well with the lack of use reported earlier. The distribution of repeat keyword is distributed from extreme to extreme and resembles a bi-modal distribution. This same type of distribution is more prominent for spell-out keyword, where the subjects are sharply divided on the essentiality of the feature. For the wallet guide, 81% of the respondents felt the guide was clustered close to very essential. Finally, background music was more uniformly distributed but skewed in the direction of being soothing music.

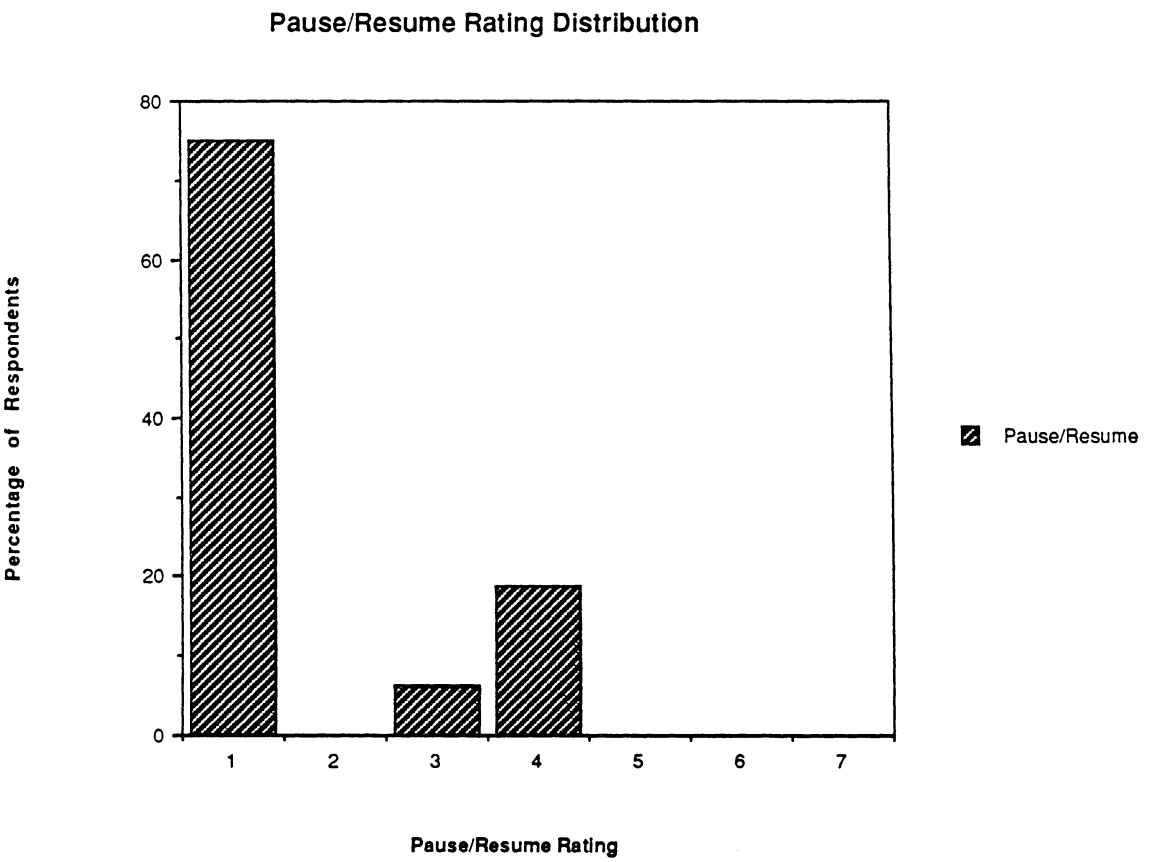


Figure 17 - Frequency Distribution for Pause/Resume Rating

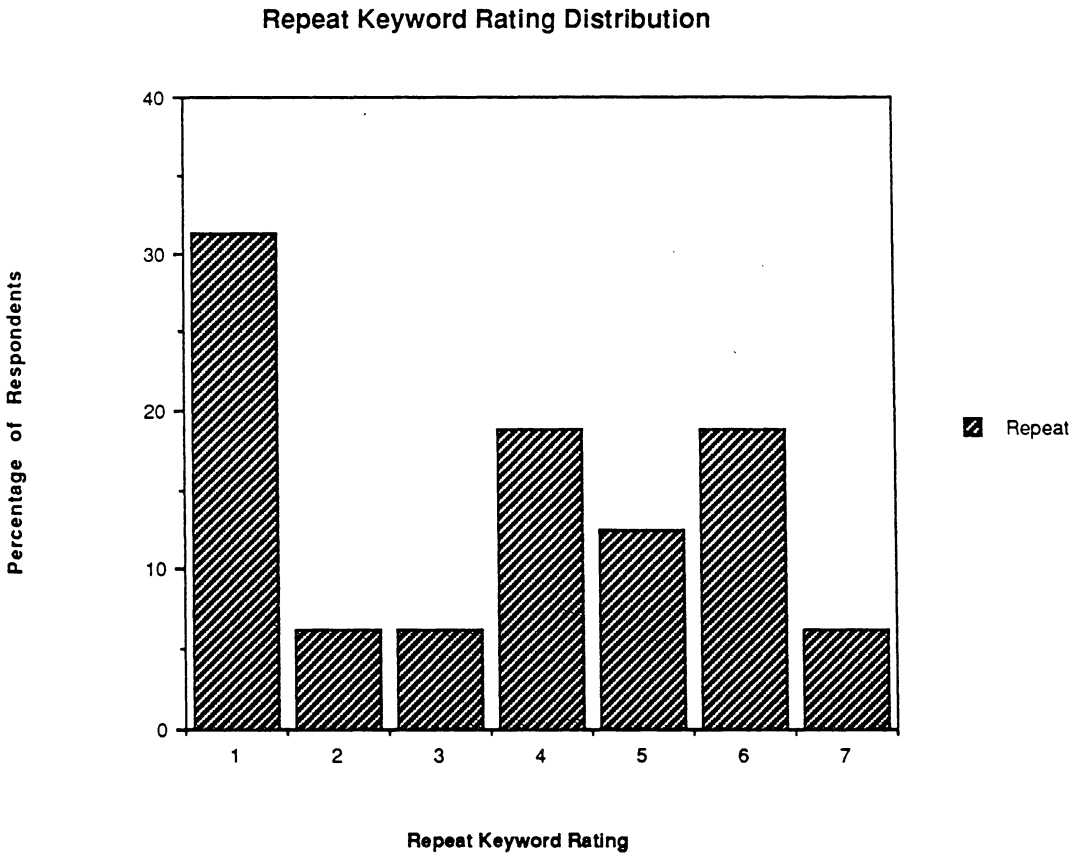


Figure 18 - Frequency Distribution for Repeat Keyword Rating

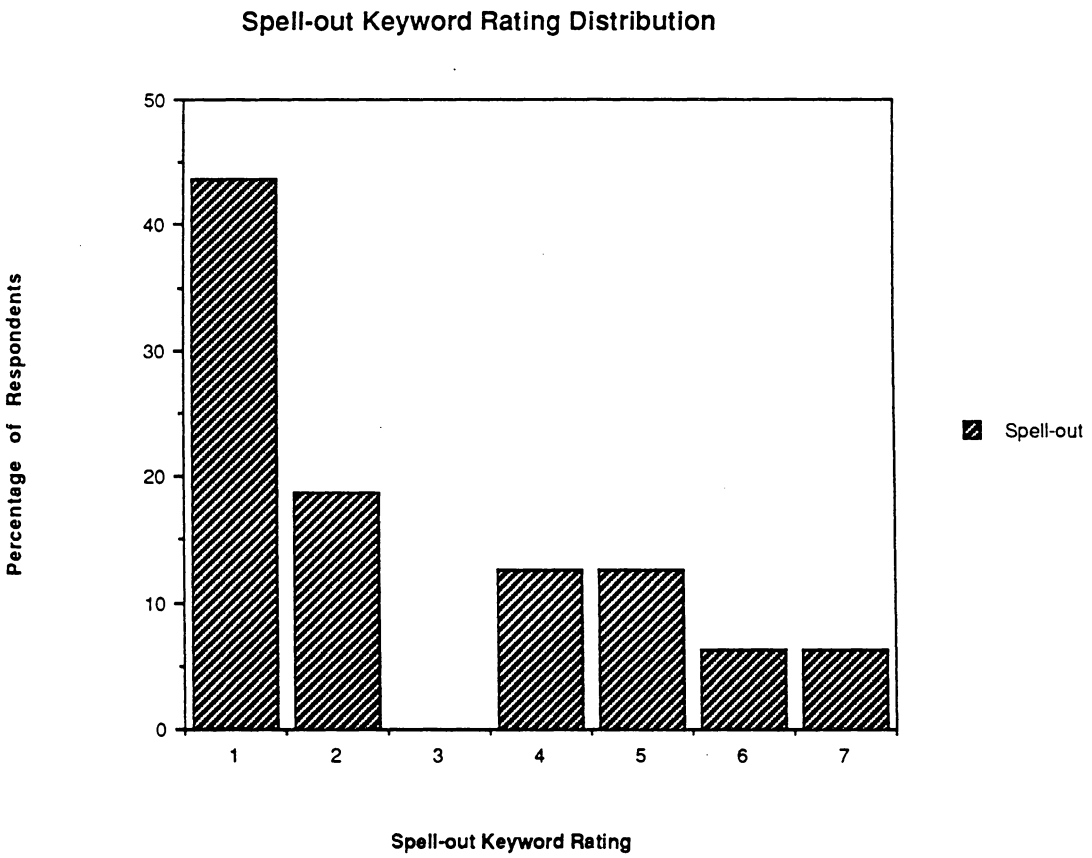


Figure 19 - Frequency Distribution for Spell-Out Keyword Rating

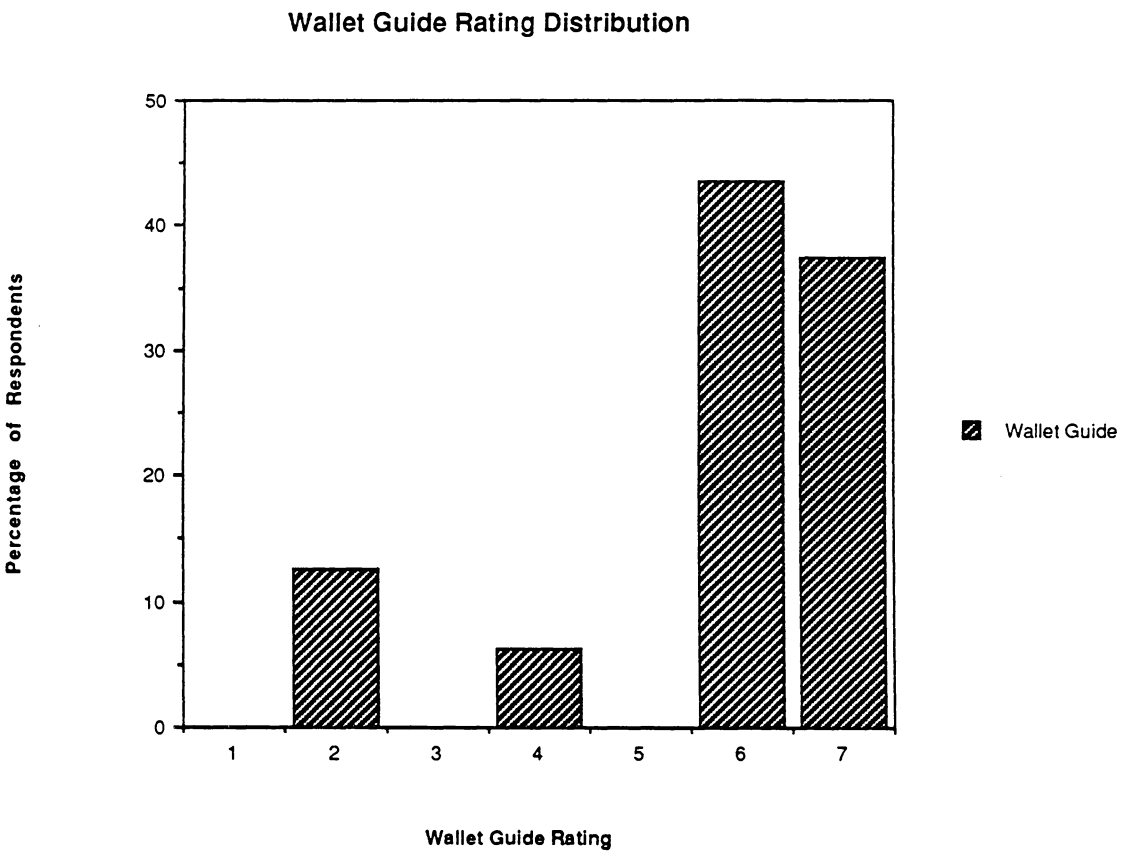


Figure 20 - Frequency Distribution for Wallet Guide Rating

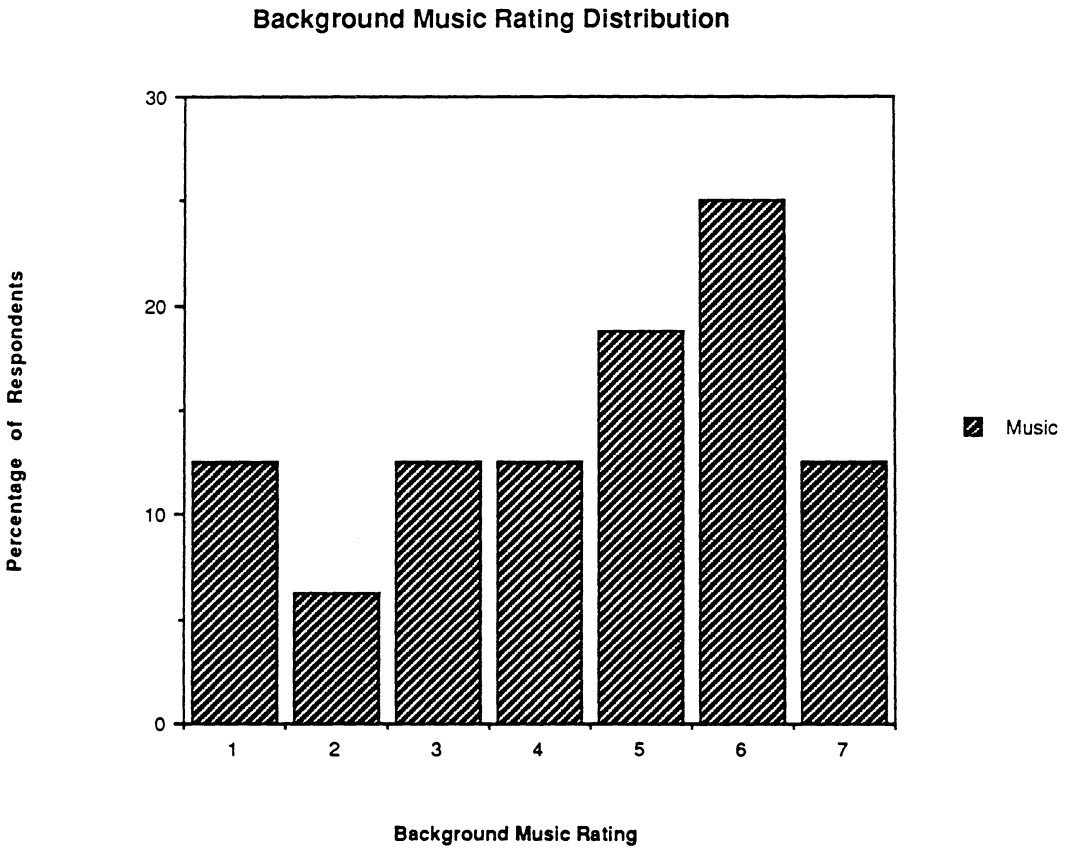


Figure 21 - Frequency Distribution for Background Music Rating

Summary of Results

Analysis of the data collected in this screening study was substantial and involved parametric (deterministic and inferential) and nonparametric testing. Table 26 summarizes the results presented earlier. Twelve out of the sixteen independent variables had a significant effect in at least one test. The four variables that had no significant effect in any test are system response time, input timeout, spell-out keyword, and selection feedback. Four additional variables only had a significant effect in the nonparametric analyses; voice type, pause/resume, repeat keyword, and command feedback. The eight remaining variables had a significant effect in the MANOVA and at least one ANOVA test. The variables are speech rate, menu organization, number of targets, wallet guide, menu feedback, background music, subject age, and subject sex. Of these eight variables, seven of them had a significant effect in at least one nonparametric test; only wallet guide did not have a significant effect on any nonparametric tests. Appendix VIII contains the reduced data obtained in this screening study.

Table 26. Summary of Results

Variable	MANOVA	ANOVA	REGRES	RATINGS
Speech Rate	*	*	*	**
Voice Type				*
Menu Organization	**	**	**	**
No. Targets	*	*	*	*
Input Timeout				
System Response Time				
Pause/Resume				*
Repeat Keyword				**
Spell-out Keyword				
Wallet Guide	*	*	*	
Selection Feedback				
Menu Feedback	*	*	*	*
Command Feedback				*
Background Music	*	*	*	*
Subject Age	**	**	**	**
Subject Sex	*	*	*	**

* significant at p<.20
** significant at p<.05

DISCUSSION

The results of this screening study are twofold. First, important design considerations for developing screening studies were identified. These considerations have been discussed earlier and involve the rationale used in selecting unconfounded independent variables, setting levels for continuous and discrete variables, determining valid and worthwhile dependent measures, and choosing an experimental design which can be manipulated as well as managed. Second, the results of this study established preliminary findings on the main effects of the 16 variables and provided direction for future research in the area of telephone information systems which use synthetic speech.

Assessing the results of this study was not a straight-forward process. With so many tests being performed a decision matrix was necessary. Those variables that were not significant in any of the tests will be held constant in future studies (4 variables). Setting of the variable levels should be based on reducing overall search time and reflect subject preferences as reflected in the appropriate subjective ratings. If a variable was significant in two or more tests, then it will be manipulated in future studies (8 variables). In the situation where a variable was significant in only one test, the variable should be evaluated in a separate paper analysis (4 variables). This analysis should involve reviewing all the available data on the variable and rendering a decision on whether or not to investigate the variable in future studies. Table 27 summarizes the disposition of the 16 variables.

Table 27. Recommendation for treatment of 16 Variables in future studies

Variables to be tested in future studies (8)

Speech Rate
Menu Organization
Number of Targets
Wallet Guide
Menu Feedback
Background Music
Subject Age
Subject Sex

Variables to be fixed at a specific level in future studies (4)

Input Timeout (2 seconds)
System Response Time (0 seconds)
Spell-out Keyword (not available)
Selection Feedback (not available)

Variables to be reviewed by paper analysis (4)

Voice Type
Pause/Resume
Repeat Keyword
Command Feedback

The results were clear regarding analysis techniques.. The deterministic ANOVA and the predictive regression analysis produced the same results. Therefore, an experimenter should choose the method that is most appropriate for the type of research being conducted. Clearly, the results of this study indicate either method could be applied. However, for bi-level testing the ANOVA is generally a more appropriate test than regression analysis. An unresolved issue, which this research did not attempt to address, is the appropriateness of the MANOVA test in a screening study. It could be argued that the MANOVA is too conservative of a test for a screening study, possibly excluding too many variables while it controls for inflated alpha across individual ANOVAs. This issue requires further analysis, because the experimenter must decide, based on the experiment, how critical it is to guard against inflating alpha when testing across several ANOVAs.

The factor analysis confirmed the trends observed in the ANOVA and regression analyses. Search Time Ratio and Search Efficiency Ratio are highly correlated, meaning that only one measure need be obtained for testing purposes. Selection of a ratio score should be dependent upon the objectives of the researcher. If a variable is to be studied because search time is of interest, than Search Time Ratio should be measured and tested. However, if the researcher is more interested in the effects of a variable on search errors, than obviously Search Efficiency Ratio is a more appropriate measure. Regardless, it should be understood that one ratio can be predicted based on the results of the other. Similarly, this applies to Strict and Synonym Message Tanscription Average. Strict Message Transcription Average should be measured if pure intelligibility is of concern to the researcher. If

perception or basic comprehension is more important, than the relaxed Synonym Message Transcription Average may be a more appropriate measure.

The frequency counts for the command features pause/resume, repeat keyword, and spell-out keyword suggest that they are not essential command features. Comparing these results with the subjective ratings for each variable suggests that pause/resume and spell-out keyword are not essential command features. Pause/resume had a significant effect in the ease-of-use rating, however, it can be argued that this single test result may be caused by a higher order interaction that is an alias with pause/resume. Spell-out keyword never had a significant effect on any parametric or nonparametric test and was used very infrequently by subjects. For repeat keyword, the subjective ratings and frequency counts suggest that the command feature may be a useful feature in the system. Additional paper analysis should be performed to decide whether this variable requires additional testing.

The subjective measures also provided a means to analyze the usability of the information system. Ratings varied from the general (e.g., ease of use) to the specific (e.g. wallet guide); thus, permitting an aggregate analysis of all the variables, as well as specific assessment of certain variables. The results indicate that the general ratings were not as sensitive to treatment conditions as were the more specific ratings. Seven of the 16 variables were significant in 32 general rating Mann-Whitney U-tests , while five of the seven variables were significant in 23 voice characteristic rating and specific variable rating Mann-Whitney U-tests.

The inability to analyze interactions statistically should not be construed as meaning that all interactions were not studied. Data gathered from experimenter

observations and debriefings provided excellent information on obvious variable interactions. As an example, having a wallet guide to read eliminated a subject's need to request that a keyword be repeated or spelled-out. Also, subjects that had wallet guides often concentrated more on reading and following the guide rather than listening to the information system. As a result, they sometimes selected too early or too late.

Subjects that had wallet guides also expressed certain frustration in not being able to "jump" or move directly to the store item. This suggests that having a spatial layout of the database causes subjects to want to activate the system, rather than react to what the system says. Future studies may want to study an information system that allows users to input, possibly type in on the telephone keypad, the desired store item rather than select keywords.

Observational data also provided important information on subjects' search strategies, design preferences, and general impressions of the information system. Some subjects felt that the most effective way to correct for an incorrect selection was to press the restart button and start over again. While many subjects used the backup key when they had selected incorrectly, there appears to be a portion of the population that does not prefer to backup or "retrace their steps."

College age subjects were unanimous in expressing opinions that they enjoyed using the system and thought it was a good application. While some subjects said that they would not personally use such a system, stating they prefer interpersonal conversations, the college age population felt that synthetic speech information systems provided an effective means of getting simple information. The middle age subjects generally agreed with the college age subjects, however,

several middle age subjects felt that while the system was entertaining to use, they would not use such a system and did not feel think that the system was a good application.

Apparantly, user demographics are an important factor that affect user performance measures as well as subjective ratings. Both subject age and sex had significant effects on objective and subjective measures. As discussed above, interview data also identified differences between age groups. These results and observations amplify the importance of studying user demographics when performing usability studies on consumer products, especially when new and innovative technology is involved.

Frequency distributions for the subjective ratings were helpful for studying those discrete variables where only half of the subjects had the treatment condition available to them (e.g. background music). The wallet guide was rated consistently as an essential feature. Conversely, pause/resume was rated consistently as an unessential feature. The benefit of having such ratings was that the frequency distributions were used to corroborate or dispute analytic testing. Clearly, this was true for pause/resume where the feature was never used, but the variable had a significant effect on the ease-of-use rating. In this situation, the test result is most likely caused by a higher order interaction which is an alias of the variable.

Subjective ratings for system response time and input timeout were as insensitive as the objective measures for these two continuous variables. Interestingly, during the experimental sessions it was readily obvious to the experimenter that the combination of the four second input timeout and four second system response time made the system very slow to use, was frustrating for some

users, and definitely induced selection errors. Unfortunately, not being able to test for variable interactions was a sacrifice in this experimental design. Because, neither system response time nor input timeout had significant effects on the dependent measures, they should be set at their minimum time levels to decrease overall search time in the system.

Both demographic variables proved to be significant variables in the screening study. However, in selecting demographic variables one should be sensitive to the feasibility of obtaining a sufficient number of subjects from the sample population. The recruitment of middle age subjects was slow and was confounded by the fact that the four subjects that failed the hearing test were middle-aged men. This age-sex bias may be worth studying in future research. Also, the subjects were all from the University community, therefore, the results of this study are not applicable to the general population. The results here and in Merkle (1988) show that demographics variables have a significant effect on user performance for telephone information systems.

Conducting the screening study was a complicated process, even with the assistance of a computerized system and data collection routine. The basic task of keeping track of settings for 16 variables required a very methodical and careful approach when setting up the software for an experimental session. Testing 16 variables also meant having at least 16 things that could be wrong in any given experimental session. Certainly, this should be a consideration whenever an experimenter is contemplating performing a screening study. Simon (1977) stated that he believed 100 variables could be tested at the same time in a human factors

screening study. As a result of conducting this study, the author has serious reservations about performing screening studies with more than 16 variables. The lesson learned being that it might well be better to conduct multiple screening studies (8 to 16 variables) than attempt to study too many variables (greater than 16) in a single experiment and risk unacceptable levels of experimenter error.

CONCLUSIONS

Foremost, this study demonstrated the worthiness of screening studies in human performance research. This study also represents a complete example and evaluation of the use of Hadamard matrices in designing and conducting efficient screening studies -- the main effects of 16 bi-level independent variables were measured with only 32 observations. Variable interactions were not evaluated in this study but will be investigated in future studies involving the remaining variables.

The lessons learned from this study also provide important guidance to researchers contemplating using screening studies in a sequential research design strategy for human factors research. In conclusion, this screening study provides the following insight:

- The results suggest that for a telephone information system which uses synthetic speech the following variables have a significant effect on objective measures of user performance: speech rate, menu organization, number of targets, wallet guide, menu feedback, background music, subject age, and subject sex. In addition, voice type, speech rate, menu organization, number of targets, repeat keyword, selection feedback, menu feedback, command feedback, background music, subject age, and subject sex had a significant effect on users' responses to subjective ratings. System response time, input timeout, pause/resume, and spell-out keyword had no effect on any dependent measure.

- Use experimental designs like Hadamard matrices or fractional factorials to perform preliminary or screening studies on a relatively large number of variables. However, keep in mind that while the experimental design may be very efficient for analyzing main effects, this is done at the cost of ignoring possible interactions between variables.

- Use a systematic analysis technique to reduce the number of variables down to a manageable level. Sixteen variables were not easily manipulated and required a sizable software program to control all the settings. However, 16 is a sizable reduction from the original 95 variables.

- Use a post-selection review to ensure that selected variables can be observed under the desired experimental procedures. When variables are not orthogonal, the effects of one variable may dominate over the effects of a related variable.

- Use discretion in selecting dependent measures. Avoid collecting data for the sake of collecting data. The metering package for this study collected and time-stamped every subject and system input/output action; thus, requiring a considerable effort to determine what measures were worth reducing from the data set. Use orthogonal objective measures, whenever possible, to avoid inflating the probability of Type I error. For subjective measures, use specific ratings to study individual variables. Specific ratings are more sensitive than general or overall ratings.

- Collect both objective and subjective data. Data collection for information systems should include both objective performance scores, as well as subjective

preference opinions. Use both sources to corroborate or refute individual findings. As an example, pause/resume was found to be significant in the Mann-Whitney U Test for the Ease-of-Use Rating. Since no subject with the ability to pause/resume a search ever chose to use the feature, the findings of the test were dismissed as an artifact of confounding the variable with other unknown interactions.

- Develop a criterion for evaluating the analysis results. Evaluate results carefully with the objective being to identify variables which have no significant effect on performance. If a variable has a definite effect on user performance, continue to study the variable in future studies. If a variable is marginal, analyze the experimental and historical data further to help support a final decision. If a variable does not have an effect on user performance, fix the variable at the most desirable level. A post-hoc analysis should be performed to determine a desirable level for a fixed variable.

REFERENCES

Allen, J., Hunnicutt, M. S., and Klatt, D. (1987). *From text to speech: The MITalk system.* Cambridge: Cambridge University Press.

Anderson, D. P. (1984). A talking computer gives weather forecasts by telephone. In *Proceedings of the 1st International Conference on Speech Technology,* (pp. 98-103). Brighton, UK: North-Holland.

Cochran, W. G., and Cox, G. M. (1957). *Experimental designs.* New York: Wiley.

Diamond, W. J. (1981). *Practical experiment designs for engineers and scientists.* Belmont, CA: Lifetime Learning.

Engel, S. E., and Granda, R. E. (1975). Guidelines for man/display interfaces (Tech. Report TR 00.2720). Poughkeepsie, NY: IBM Poughkeepsie Laboratory.

Gould, J. D., and Boies, S. J. (1984). Speech filing - An office system for principles. *IBM Systems Journal, 23*/1, 65-81.

Greene, B. G., Manous, L. M., and Pisoni, D. B. (1984). Perceptual evaluation the DECtalk: a final report of version 1.8. In *Research on Speech Perception Report No 10* (pp. 77-127). Bloomington, IN: Indiana University.

Hise, H. H., and Lundin, F. J. (1985). Text-to-speech quality in a telephone information system. *Journal of the American Voice I/O Society, 2*, 65-74.

Kidd, A. L. (1982). Problems in man-machine dialogue design. In *IEEE Proceedings of Sixth Conference on Computer Communications* (pp. 531-536). North-Holland, Amsterdam: North-Holland.

Kiger, J. I. (1984). The depth/breadth trade-off in the design of menu-driven user interfaces. *International Journal of Man-Machine Studies, 20*, 201-213.

Logan, J. S., Pisoni, D. B., and Greene, B. G. (1985). Measuring the segmental intelligibility of synthetic speech: Results from eight text-to-speech systems. In *Research on Speech Perception Progress Report No 11* (pp. 3-32). Bloomington, IN: Indiana University.

Manous, L. M., Pisoni, D. B., Dedina, M. J., and Nusbaum, H.C. (1985). Comprehension of natural and synthetic speech using a sentence verification task. In *Research on Speech Perception Report No. 11* (pp. 33-57). Bloomington, IN: Indiana University.

Merkle, P. J. (1988). *Using subjective ratings to select independent variables in equipment design research: A validation study.* Unpublished masters thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Merva, M. A. (1987). *The effects of speech rate, message repetition, and information placement on synthesized speech intelligibility.* Unpublished masters thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Munger, S., Smith, R. W., and Payne, D. (1962). *An index of electronic equipment operability: Data store.* Pittsburgh: American Institutes for Research.

Nakatani, L. H., and O'Connor, K. D. (1980). Speech feedbacking for touch-keying. *Ergonomics, 23,* 643-654.

Podgorny, P. (1985). Telephone as computer terminal. In *The Official Proceedings of Speech Tech `85* (pp. 103-109). New York: Media Dimensions.

Rosson, M. B., and Cecala, A. J. (1985). *Designing a quality voice: An analysis of listener's reactions to synthetic voices* (Research Report RC11398). Yorktown Heights, NY: IBM Watson Research Center.

Rosson, M. B., and Mellen, N. M. (1985). *Behavioral issues in speech-based remote information retrieval* (Research Report RC11028). Yorktown Heights, NY: IBM Watson Research Center.

Sanders, M. S., and McCormick, E. J. (1987). *Human factors in engineering design.* New York: McGraw-Hill.

SAS Institute Inc. (1986). *Statistical analysis system, release 5.16.* Cary, NC: SAS Institute Inc.

Schmandt, C. (1985a). Voice access to an electronic mail system. In *The Official Proceedings of Speech Tech `85* (pp. 89-91). New York: Media Publications.

Schmandt, C. (1985b). Voice communications with computers. In H. R. Hartson (Ed.), Advances in Human-Computer Interaction, Volume I (pp. 133-159). Norwood, NJ: Ablex.

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

Simon, C. W. (1977). *Design, analysis, and interpretation of screening designs for human factors engineering research* (Technical Report CWS-03-77A). Westlake, CA: Canyon Research Group. (AD 056-985).

Simon, C. W., and Roscoe, S. N. (1984). Application of a multifactor approach to transfer of training research. *Human Factors, 26,* 591-612.

Simpson, C. A., and Marchionda-Frost, K. (1984). Synthesized speech rate and pitch effects on intelligibility of warning messages for pilots. *Human Factors, 26,* 509-517.

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*, 115-141.

Slowiaczek, L. M., and Nusbaum, H. C. (1985). Effects of speech rate and pitch contour on the perception synthetic speech. *Human Factors, 27*, 701-712.

Tatro, J. S., and Roscoe, S. N. (1986). An integrated display for vertical and translational flight: Eight factors affecting pilot performance. *Human Factors, 28*, 101-120.

Thomas, J C., Rosson, M B., and Chodorow, M. (1984). Human factors and synthetic speech. In *Proceedings of the Human Factors Society 28th Annual Meeting* (pp. 763-767). Santa Monica, CA: Human Factors Society.

Waterworth, J., and Lo. A. (1984). Examples of an experiment: Evaluating some speech synthesizers for public announcements. In A. Monk (Ed.), *Fundamentals of human-computer interaction.* London: Academic Press.

Whitehurst, , H. O. (1982). Screening designs used to estimate the relative effects of display factors on dial reading. *Human Factors, 24*, 301-310.

Witten, I. H., and Mandams, P. H. C. (1977). The telephone inquiry service: A man-machine system using synthetic speech. *International Journal of Man-Machine Studies, 9*, 449-464.

Appendix I

Informed Consent Form

Participant's Informed Consent Form

The following experiment is a study concerning the evaluation of a telephone based information system. During the experiment, you will be monitored with a closed circuit video system. As a participant in this experiment, you have certain rights as explained below. The purpose of this document is to describe these rights and to obtain your written consent to participate in the experiment.

1. You have the right to discontinue your participation in the study at any time for any reason. If you decide to terminate the experiment, inform the researcher and he will pay you for the length of time you have participated.

2. You have the right to inspect your data and withdraw it from the experiment if you feel that you should for any reason. In general, data are processed and analyzed after a subject has completed the experiment. At that time, all identification information will be removed and the data treated with anonymity. Therefore, if you wish to withdraw your data, you must do so immediately after your participation is completed.

3. You have the right to be informed of the overall results of the experiment. If you wish to receive a synopsis of the results, include your address with your signature below. If after receiving the synopsis, you would like more indepth information, please contact Virginia Tech's Human-Computer Interaction Laboratory and a full report will be made available to you.

This research is funded by a research contract with the National Science Foundation. The co-principal investigators are Dr. Robert Williges, and Ms. Beverly Williges. The researcher is Douglas B. Beaudet. All of these people can be contacted at the following address and phone number:

Human-Computer Interaction Laboratory
302 Whittemore Hall
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

Further comments or questions can be addressed to Charles Waring, chairman of the Institutional Review Board for the Use of Human Subjects in Research. He can be contacted at the address and the phone number listed below:

Mr. Charles Waring
Office of Sponsored Research Programs
301 Burruss Hall
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

If you have any questions about the experiment or your rights as a participant, please do not hesitate to ask. The researcher will do his best to answer them, subject only to the constraint that he does not pre-bias the experimental results.

Your signature below indicates that you have read and understand your rights as a participant (as stated above), and that you consent to participate.

_____

Participant's Signature

_____

Witness's Signature

_____

_____

_____

Print name and address if you wish to receive a summary of the experimental results.

Appendix II

Pre-test Questionnaire

Subject number _____

Subject Information Questionnaire

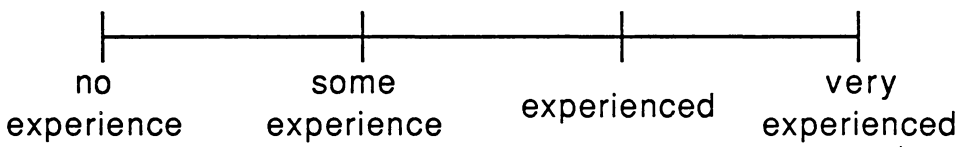
Age: _____ Sex: _____ Native language: _____
(include region of rearing)

Please list any hearing impairments you may have.

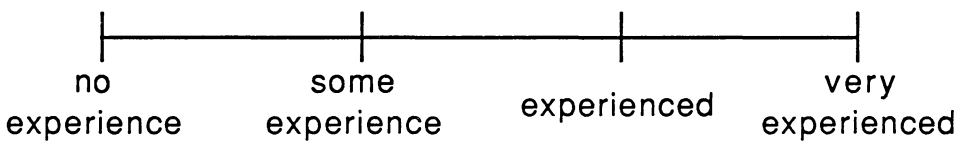
_____

For the following questions, please circle the most accurate response.

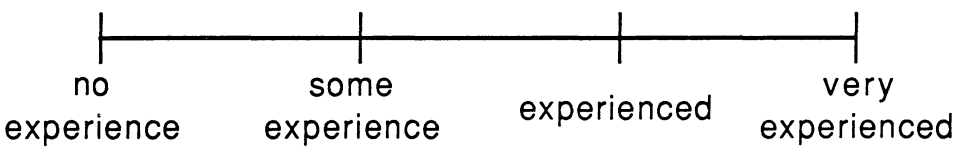
- How experienced are you with using computers?



- How experienced are you with using information systems?



- How experienced are you with listening to synthesized speech?



Appendix III

Subject's Instructions

Single Target Instructions

Your task is to search for information on store items in the department store's talking database. Store items will be presented as targets on the computer display in front of you. You will find the target by using the telephone keys to move through the talking database.

These are your instructions:

1. Press the ON/OFF key on the telephone keypad and listen for a dialtone.
2. Press the DIAL key on the telephone keypad (upper right corner).
3. The talking computer will answer the telephone and offer you instructions. Press the **# key** on the telephone keypad and listen carefully to the instructions for using the telephone keypad.
4. Read the first target on the computer display in front of you.
5. Watch the computer display. It will signal you when the search is about to begin.
6. The talking computer will begin speaking a menu of keywords.Keywords categorize groups of store items. After each keyword is spoken, the computer will pause briefly to allow you to select the item.If you do not select the item, the computer will speak another keyword for that menu.
7. To locate the target, select a keyword from the menu which best categorizes the store item you are searching for. The computer will then speak a new menu of keywords, based on your selection. If you need to hear the keypad instructions again, select HELP from any menu.
8. Continue listening to menus and selecting keywords until you reach the desired store item.
9. When you hear the desired store item, press the **2 key** on the telephone keypad and listen carefully to the information message.

10. The computer display will prompt you to transcribe what you heard.
11. Type the information message you heard into the computer, and press the RETURN key.
12. Rate the certainty of your transcription being correct on a scale of 1 (very uncertain) to 7 (very certain), and press the RETURN key.
13. Rate the difficulty of understanding the message on a scale of 1 (very difficult) to 7 (very easy), and press the RETURN key.
14. Rate the difficulty of locating the store item on a scale of 1 (very difficult) to 7 (very easy), and press the RETURN key.
15. Read the next target on the computer display and get ready to start the next search. The computer display will signal you to begin the next search and will speak the first item in the main menu. Locate the next target and transcribe the information message.
16. The experiment will proceed in this fashion. You will search for a total of 16 targets.
17. The computer display will indicate when you have completed the target searches. The computer display will then request that you rate certain characteristics of the telephone information system. The meaning of each characteristic and how it should be rated will be explained on the computer display.

If you have any questions, please ask the experimenter now.

Multiple Target Instructions

Your task is to search for information on store items in the department store's talking database. Two store items will be presented as a target on the computer display in front of you. You will find each store item by using the telephone keys to move through the talking database.

These are your instructions:

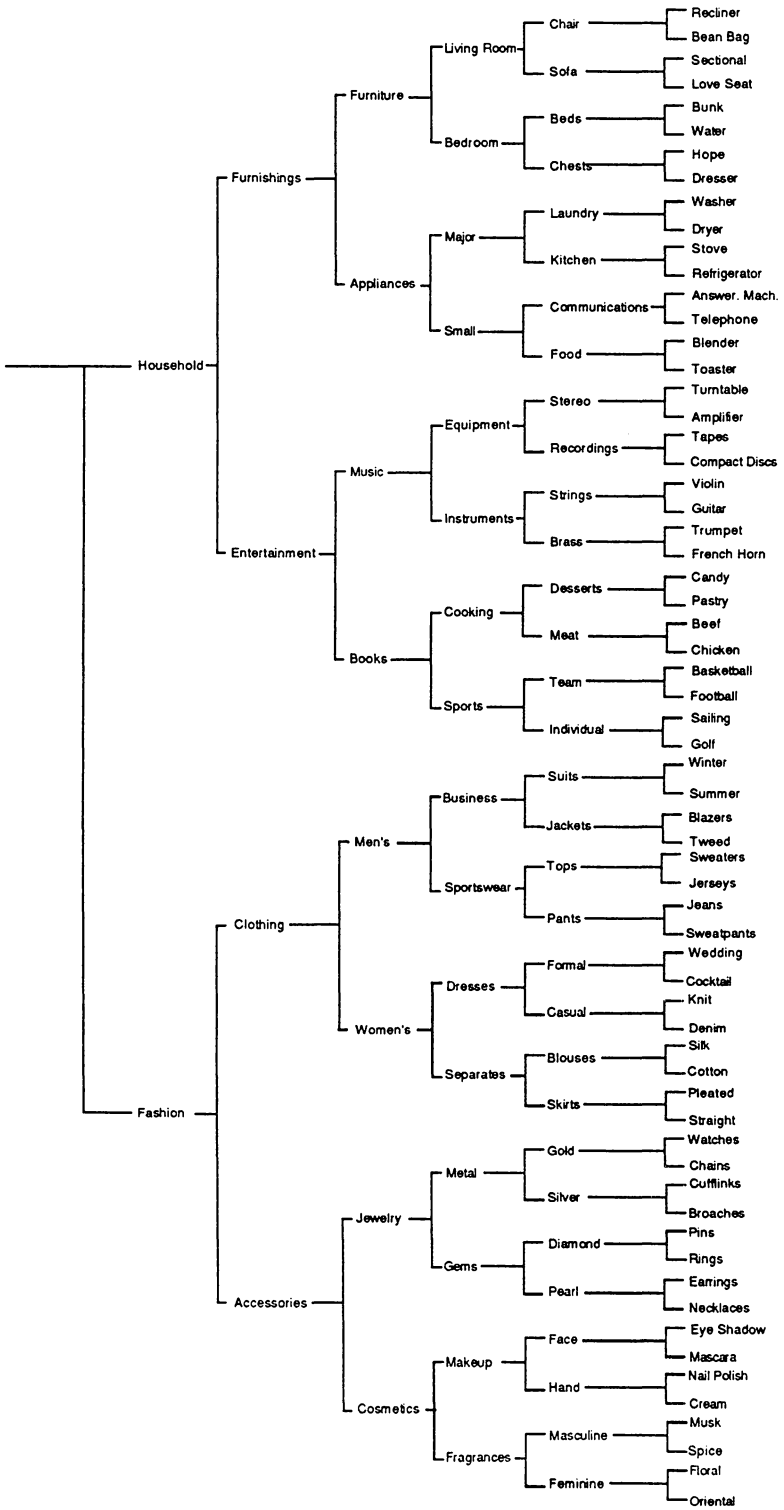
1. Press the ON/OFF key on the telephone keypad and listen for a dial tone.
2. Press the DIAL key on the telephone keypad (upper right corner).
3. The talking computer will answer the telephone and offer you instructions. Press the **# key** on the telephone keypad and listen carefully to the instructions for using the telephone keypad.
4. Read the first target on the computer display in front of you. The target requests that you find the information message for two store items.
5. Watch the computer display. It will signal you when the search is about to begin. Begin by searching for the first store item.
6. The talking computer will begin speaking a menu of keywords. Keywords categorize a group of store items. After each keyword is spoken, the computer will pause briefly to allow you to select the item. If you do not select the item, the computer will speak another keyword for that menu.
7. To locate the first store item, select a keyword from the menu which best categorizes the store item you are searching for. The computer will then speak a new menu of keywords, based on your selection. If you need to hear the keypad instructions again, select HELP from any menu.
8. Continue listening to menus and selecting keywords until you reach the desired store item.
9. When you hear the desired store item, press the **2 key** on the telephone keypad and listen carefully to the information message.
10. The computer display will prompt you to transcribe what you heard.
11. Type the information message you heard into the computer, and press the RETURN key on the computer keyboard.

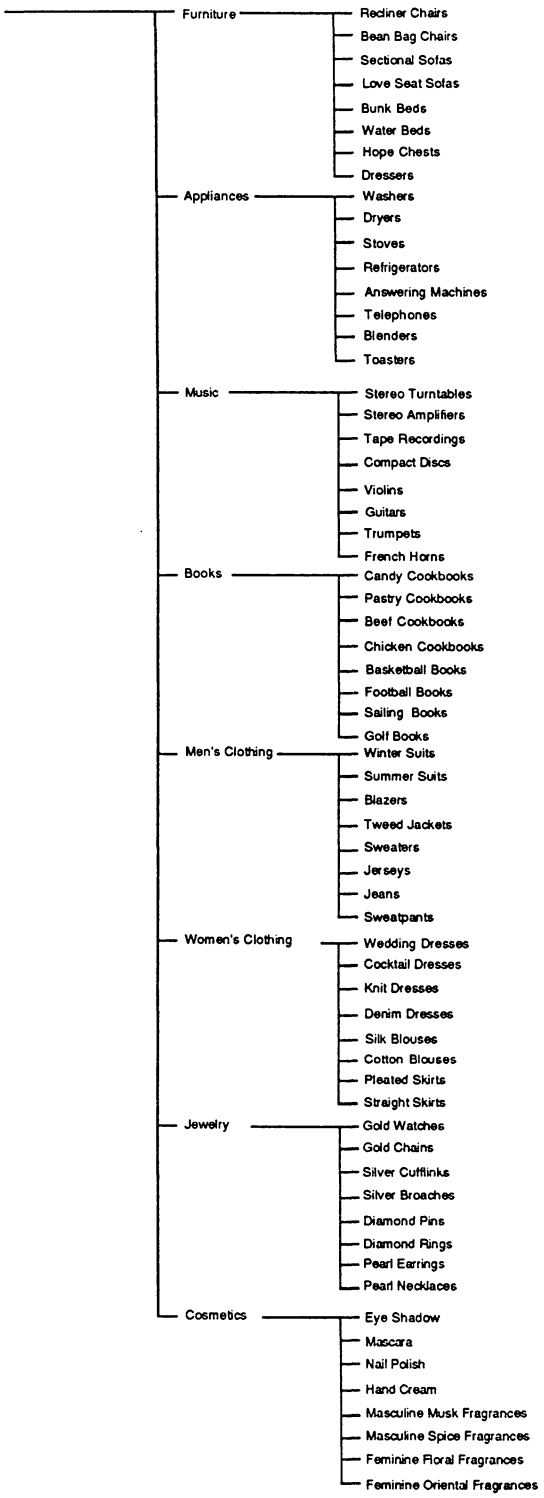
12. Rate the certainty of your transcription being correct on a scale of 1 (very uncertain) to 7 (very certain), and press the RETURN key.
13. Rate the difficulty of understanding the message on a scale of 1 (very difficult) to 7 (very easy), and press the RETURN key.
14. Rate the difficulty of locating the store item on a scale of 1 (very difficult) to 7 (very easy), and press the RETURN key.
15. Review the target on the computer display and get ready to start the search for the second store item. The computer display will signal you to begin the search and will speak the information message for the first store item.
16. Search for the second store item using the same techniques which were described in steps 7 through 10. **Remember**, you are starting the search from the information message for the first store item, not from the main menu.
17. Locate the second store item and transcribe the information message.
18. Read the next target on the computer display and get ready to start the next search. The computer display will signal you to begin the next search and will speak the first item in the main menu. Locate each store item and transcribe their respective information messages.
19. The experiment will proceed in this fashion. You will search for a total of 16 targets.
20. The computer display will indicate when you have completed the target searches. The computer will then request that you rate certain characteristics of the telephone information system. The meaning of each rating and how it should be rated will be explained on the computer display.

If you have any questions, please ask the experimenter now.

Appendix IV

Information Databases


Household
Fashion
Furnishings
Entertainment
Clothing
Accessories
Furniture
Appliances
Music
Books
Men's
Women's
Living Room
Bedroom
Major
Small
Equipment
Instruments
Cooking
Sports
Business
Sportswear
Dresses
Separates
Jewelry
Cosmetics
Chair
Sofa
Beds
Chests
Laundry
Kitchen
Communications
Food
Stereo
Recordings
Strings
Brass
Desserts
Meat
Team
Individual
Suits
Jackets
Tops
Pants
Formal
Casual
Blouses
Skirts
Metal
Gems
Makeup
Fragrances
Gold
Silver
Diamond
Pearl
Face
Hand
Masculine
Feminine
Recliner
Bean Bag
Sectional
Love Seat
Bunk
Water
Hope
Dresser
Washer
Dryer
Stove
Refrigerator
Answer. Mach.
Telephone
Blender
Toaster
Turntable
Amplifier
Tapes
Compact Discs
Violin
Guitar
Trumpet
French Horn
Candy
Pastry
Beef
Chicken
Basketball
Football
Sailing
Golf
Winter
Summer
Blazers
Tweed
Sweaters
Jerseys
Jeans
Sweatpants
Wedding
Cocktail
Knit
Denim
Silk
Cotton
Pleated
Straight
Watches
Chains
Cufflinks
Broaches
Pins
Rings
Earrings
Necklaces
Eye Shadow
Mascara
Nail Polish
Cream
Musk
Spice
Floral
Oriental


Furniture
Recliner Chairs
Bean Bag Chairs
Sectional Sofas
Love Seat Sofas
Bunk Beds
Water Beds
Hope Chests
Dressers
Appliances
Washers
Dryers
Stoves
Refrigerators
Answering Machines
Telephones
Blenders
Toasters
Music
Stereo Turntables
Stereo Amplifiers
Tape Recordings
Compact Discs
Violins
Guitars
Trumpets
French Horns
Books
Candy Cookbooks
Pastry Cookbooks
Beef Cookbooks
Chicken Cookbooks
Basketball Books
Football Books
Sailing Books
Golf Books
Men's Clothing
Winter Suits
Summer Suits
Blazers
Tweed Jackets
Sweaters
Jerseys
Jeans
Sweatpants
Women's Clothing
Wedding Dresses
Cocktail Dresses
Knit Dresses
Denim Dresses
Silk Blouses
Cotton Blouses
Pleated Skirts
Straight Skirts
Jewelry
Gold Watches
Gold Chains
Silver Cufflinks
Silver Broaches
Diamond Pins
Diamond Rings
Pearl Earrings
Pearl Necklaces
Cosmetics
Eye Shadow
Mascara
Nail Polish
Hand Cream
Masculine Musk Fragrances
Masculine Spice Fragrances
Feminine Floral Fragrances
Feminine Oriental Fragrances

Appendix V

Targets and Information Messages

Store Items for Single Target Searches

The following are the single target store items and associated information messages. Information messages are classified as being an availability (A), information (I), location (L), or price (P) oriented messages.

Recliner Chairs: Leather coverings are offered to wholesale buyers. (I)

Hope Chests: Walnut stains are reduced by 34 to 40%. (P)

Washers: Deluxe models are available with green trimming. (A)

Food Blenders: Boxes and cartons are in the wrapping center. (L)

Guitars: Carrying cases are reduced by 55 to 63%. (P)

Compact Discs: Head cleaners are on aisle 12. (L)

Chicken Cookbooks: Collector editions are available in limited quantities. (A)

Football Books: Faculty discounts are offered to gym teachers. (I)

Men's Blazers: Garment bags are offered with new purchases. (I)

Men's Sweaters: Rugby letters are sold for $11.60. (P)

Knit Dresses: Designer collections are available in red and ivory. (A)

Silk Blouses: Maternity wear is near ladies lingerie. (L)

Gold Chains: Instant financing is available at the central office. (A)

Pearl Necklaces: Sorority clasps are in the school department. (L)

Eye Mascara: Travel supplies are sold for $17.50. (P)

Oriental Fragrances: Manufacturer's samplers are offered to interested shoppers.(I)

Store Items for the Multiple Target Searches

The following are the first store items and associated information messages for the multiple target s . The second store item are the single store items presented above. Information messages are classified as being an availability (A), information (I), location (L), or price (P) oriented messages.

Love Seat Couches: Many patterns are on the west wall. (L)

Water Beds: Merchandise exchanges are required within 1 month. (I)

Refrigerators: Large freezers are sold for $500.65. (P)

Telephones: Mouth pieces are available at the cash register. (A)

Stereo Turntables: Rubber belts are sold for $6.49. (P)

Trumpets: Free demonstrations are offered at most locations.(I)

Candy Cookbooks: Holiday specials are available in hardcover and paperback. (A)

Sailing Books: Instruction manuals are near the road maps. (L)

Men's Summer Suits: Seer sucker cloth is available in 4 colors. (A)

Men's Sweatpants: Running gear is near the camping tents. (L)

Cocktail Dresses: Colonial styles are reduced by 35 to 45%. (P)

Straight Skirts: Spring mini-skirts are offered with matching pocketbooks. (I)

Silver Cufflinks: Antique objects are sold for $64.99. (P)

Diamond Pins: Custom orders are offered to graduate students. (I)

Nail Polish: Applicator brushes are near the hair shampoo. (L)

Musk Fragrances: Bottled after-shave is available with rebate coupons. (A)

Appendix VI

Subjective Rating Scales

Subjective ratings collected for each store item and information message transcription.

Certainty in Transcribing Information Message

1-----2-----3-----4-----5-----6-----7
very very
uncertain certain

On a scale of 1 to 7, how CERTAIN are you of your transcription?

Difficulty of Understanding Information Message

1-----2-----3-----4-----5-----6-----7
very very
difficult easy

On a scale of 1 to 7, how DIFFICULT was it to understand the message?

Difficulty of Locating Store Item

1-----2-----3-----4-----5-----6-----7
very very
difficult easy

On a scale of 1 to 7, how DIFFICULT was it to locate the store item?

Subjective ratings collected after all the target searches are completed.

Ease-of-use

1-----2-----3-----4-----5-----6-----7
very very
difficult easy

On a scale of 1 to 7, how easy was the information system to use?

Computer Voice Intelligibility

1-----2-----3-----4-----5-----6-----7
very very
unintelligible intelligible

On a scale of 1 to 7, how intelligible was the computer voice?

Computer Voice Naturalness

1-----2-----3-----4-----5-----6-----7
very very
unnatural natural

On a scale of 1 to 7, how natural was the computer voice?

Computer Voice Speech Rate

1-----2-----3-----4-----5-----6-----7
very very
slow fast

On a scale of 1 to 7, how fast did the computer voice speak?

System Response Time

1-----2-----3-----4-----5-----6-----7
very very
slow fast

On a scale of 1 to 7, how quickly did the information system respond to your actions?

Input Timeout

1-----2-----3-----4-----5-----6-----7
very very
little much

On a scale of 1 to 7, how much time did you have to enter a command?

Database Organization

1-----2-----3-----4-----5-----6-----7
very very
complex simple

On a scale of 1 to 7, how complex was the organization of the database?

Pause/Resume Speech

1-----2-----3-----4-----5-----6-----7
not absolutely
essential essential

On a scale of 1 to 7, how essential was the pause/resume feature?

Repeat Keyword

1-----2-----3-----4-----5-----6-----7
not absolutely
essential essential

On a scale of 1 to 7, how essential was the repeat keyword feature?

Spell-out Keyword

1-----2-----3-----4-----5-----6-----7
not absolutely
essential essential

On a scale of 1 to 7, how essential was the spell-out keyword feature?

Wallet Guide

1-----2-----3-----4-----5-----6-----7
not absolutely
essential essential

On a scale of 1 to 7, how essential was the wallet guide feature?

Background Music

1-----2-----3-----4-----5-----6-----7
very very
disruptive soothing

On a scale of 1 to 7, what did you think of the background music?

Appendix VII

Mann-Whitney U Test Table

Summary of Results from Mann-Whitney U Tests

RATING (Variable)	U-Value*
GENERAL RATINGS (Variables: All 16 Variables)	
Store Item Search Difficulty Rating	
Type of Voice	122
Speech Rate	128
Menu Organization	112
Number of Targets	128
Input Timeout	106
System Response Time	100
Pause/Resume	128
Repeat Keyword	106
Spell-out Keyword	116
Wallet Guide	116
Background Music	94
Selection Feedback	106
Menu Feedback	94
Command Feedback	78
Subject Age	106
Subject Sex	116
Ease-of-Use Rating	
Type of Voice	122
Speech Rate	128
Menu Organization	87.5
Number of Targets	86
Input Timeout	106
System Response Time	100
Pause/Resume	93
Repeat Keyword	106
Spell-out Keyword	116
Wallet Guide	116
Background Music	94
Selection Feedback	106
Menu Feedback	88
Command Feedback	78
Subject Age	80
Subject Sex	116

Summary of Results from Mann-Whitney U Tests - Continued

RATING (Variable)	U-Value*
VOICE CHARACTERISTIC RATINGS	
(Variables: Voice Type, Speech Rate, Subject Age, and Subject Sex)	
Message Transcription Certainty Rating	
Voice Type	85
Speech Rate	128
Subject Age	120
Subject Sex	120
Message Transcription Difficulty Rating	
Voice Type	91.5
Speech Rate	110.5
Subject Age	109
Subject Sex	119.5
Computer Voice Intelligibility Rating	
Voice Type	125.5
Speech Rate	121.5
Subject Age	55
Subject Sex	112
Computer Voice Naturalness Rating	
Voice Type	90.5
Speech Rate	123.5
Subject Age	81.5
Subject Sex	124
Computer Voice Speech Rate Rating	
Voice Type	126
Speech Rate	65
Subject Age	75.5
Subject Sex	73.5

Summary of Results from Mann-Whitney U Tests - Continued

RATING (Variable)	U-Value*
VARIABLE SPECIFIC RATINGS	
(Variables: Menu Organization, System Response Time, User Input Timeout)	
Menu Organization Rating	
Menu Organization	64.5
System Response Time Rating	
System Response Time	104
User Input Timeout Rating	
Input Timeout	115

* Test significant at $p<.20$ when $U<95$

Appendix VIII

Screening Study Data

	Subjects	Condition	Voice Type	Speech Rate	Dbase Org	Targets
1	1	2	paul	Two-Forty	Eight	One
2	2	6	betty	Two-Forty	Two	Two
3	3	9	paul	Two-Forty	Eight	One
4	4	11	paul	One-Eighty	Two	Two
5	5	13	betty	One-Eighty	Two	One
6	6	22	betty	Two-Forty	Eight	Two
7	7	27	paul	One-Eighty	Two	Two
8	8	12	betty	Two-Forty	Eight	Two
9	9	16	betty	One-Eighty	Two	One
10	10	3	paul	Two-Forty	Two	One
11	11	1	paul	One-Eighty	Eight	One
12	12	26	betty	Two-Forty	Two	Two
13	13	19	paul	One-Eighty	Eight	One
14	14	25	paul	Two-Forty	Two	One
15	15	32	betty	One-Eighty	Eight	One
16	16	14	paul	One-Eighty	Eight	Two
17	17	18	betty	One-Eighty	Eight	One
18	19	28	paul	Two-Forty	Eight	Two
19	20	5	paul	Two-Forty	Two	Two
20	21	24	paul	Two-Forty	Eight	Two
21	22	21	paul	One-Eighty	Two	One
22	24	31	betty	One-Eighty	Eight	Two
23	25	8	paul	One-Eighty	Eight	Two
24	26	29	betty	Two-Forty	Two	One
25	28	30	betty	One-Eighty	Two	Two
26	30	17	betty	One-Eighty	Eight	Two
27	31	20	betty	Two-Forty	Eight	One
28	32	15	betty	Two-Forty	Eight	One
29	33	10	betty	Two-Forty	Two	One
30	34	23	paul	One-Eighty	Two	One
31	36	4	paul	Two-Forty	Two	Two
32	37	7	betty	One-Eighty	Two	Two

	Sys. Resp. Time	Input Timeout	Pause/Resume	Repeat Keyword	Spell Keyword	Wallet Guide
1	Zero	Four	A	NA	A	NA
2	Four	Two	A	NA	A	A
3	Four	Four	NA	NA	A	A
4	Zero	Four	A	NA	NA	NA
5	Four	Four	A	A	A	NA
6	Zero	Four	NA	A	NA	A
7	Four	Four	NA	NA	NA	A
8	Four	Four	A	A	NA	NA
9	Zero	Four	NA	A	A	A
10	Zero	Two	A	A	NA	A
11	Four	Four	A	A	NA	A
12	Zero	Two	NA	NA	A	NA
13	Zero	Four	NA	A	NA	NA
14	Four	Two	NA	A	NA	NA
15	Zero	Two	NA	NA	NA	NA
16	Zero	Two	A	A	A	A
17	Four	Two	A	NA	NA	A
18	Four	Two	A	NA	NA	NA
19	Four	Four	A	A	A	A
20	Zero	Two	NA	NA	NA	A
21	Zero	Two	NA	NA	A	A
22	Four	Four	NA	NA	A	NA
23	Four	Two	NA	A	A	NA
24	Four	Four	NA	NA	NA	NA
25	Zero	Two	A	A	NA	NA
26	Zero	Four	A	NA	A	A
27	Zero	Two	A	A	A	NA
28	Four	Two	NA	A	A	A
29	Zero	Four	A	NA	NA	A
30	Four	Two	A	NA	A	NA
31	Zero	Four	NA	A	A	NA
32	Four	Two	NA	A	NA	A

	Selection F-back	Menu F-Back	Comand F-back	Back. Music	Age	Sex
1	NA	NA	A	Music	CA	F
2	NA	NA	NA	None	CA	F
3	A	A	NA	None	CA	M
4	A	A	NA	None	CA	F
5	NA	A	NA	Music	CA	M
6	NA	A	NA	Music	CA	F
7	NA	NA	A	Music	CA	M
8	A	NA	A	None	CA	M
9	A	NA	A	None	CA	F
10	A	NA	NA	Music	CA	M
11	NA	NA	NA	None	MA	F
12	A	A	A	Music	CA	M
13	A	A	A	Music	MA	M
14	NA	A	A	None	CA	F
15	NA	NA	NA	None	CA	M
16	NA	A	A	None	CA	M
17	A	A	A	Music	CA	F
18	NA	A	NA	Music	MA	M
19	A	A	A	Music	MA	F
20	A	NA	A	None	MA	F
21	NA	A	NA	Music	MA	F
22	NA	A	A	None	MA	F
23	A	NA	NA	Music	CA	F
24	A	NA	NA	Music	MA	F
25	NA	NA	A	Music	MA	F
26	A	NA	NA	Music	MA	M
27	A	A	NA	None	MA	F
28	NA	NA	A	Music	MA	M
29	NA	A	A	None	MA	M
30	A	NA	A	None	MA	M
31	NA	NA	NA	None	MA	M
32	A	A	NA	None	MA	M

	Search Time Ratio	Search Efficiency Ratio	Invalid Keypress Average	Strict Trans. Average	Synonym Trans. Average
1	.93807	.93506	.12500	3.43750	3.50000
2	.87734	.98494	0	2.96875	3.09375
3	.97612	1.00000	.06250	3.50000	3.62500
4	.72485	.82785	.06250	2.87500	3.00000
5	.78414	.83237	0	3.43750	3.62500
6	.83140	.86275	.34375	3.43750	3.53125
7	.97215	1.00000	0	3.65625	3.68750
8	.91707	.99676	.06250	3.56250	3.56250
9	1.08845	1.00000	0	3.62500	3.93750
10	1.01661	.97959	0	3.43750	3.50000
11	1.01006	1.00000	.06250	3.81250	3.81250
12	.61612	.74318	.12500	2.37500	2.56250
13	.50778	.65455	.43750	3.31250	3.56250
14	.61015	.69903	0	3.25000	3.37500
15	1.01150	1.00000	0	3.87500	3.93750
16	.93070	.96250	.06250	3.71875	3.78125
17	.94461	.92903	0	3.87500	3.87500
18	.80804	.88000	.59375	3.21875	3.37500
19	.75089	.70172	.18750	2.84375	3.03125
20	.66929	.76427	.31250	3.21875	3.37500
21	1.04721	.97959	0	2.62500	2.87500
22	.83834	.93333	.03125	2.43750	2.65625
23	.89905	.94479	.09375	3.46875	3.50000
24	.45881	.57600	.06250	2.31250	2.50000
25	.83626	.85827	.03125	3.40625	3.50000
26	.83650	.88252	.31250	3.65625	3.75000
27	.70907	.79121	0	2.50000	2.81250
28	.82752	.84211	0	2.81250	3.25000
29	.95910	.97959	0	3.06250	3.18750
30	.54837	.59504	0	3.43750	3.50000
31	.89464	.94236	0	3.65625	3.68750
32	.75426	.81343	.18750	3.56250	3.59375

	Certainty Rating	Difficulty Rating	Difficulty Search Rating	Ease-of-Use	Intelligibility	Naturalness
1	5	5	7	6	5	5
2	5	5	7	6	5	5
3	6	5	7	6	5	3
4	7	5	7	6	5	4
5	1	3	7	5	4	4
6	3	3	5	5	5	4
7	3	5	6	6	6	5
8	5	6	7	6	5	3
9	7	6	6	4	5	3
10	7	6	7	6	5	4
11	6	6	7	7	5	4
12	2	3	4	5	5	4
13	4	3	6	4	3	4
14	6	5	7	5	4	5
15	3	4	7	6	5	5
16	6	6	7	5	5	6
17	3	3	5	6	3	2
18	5	5	6	6	4	3
19	6	6	6	4	4	5
20	2	2	6	5	2	2
21	3	3	6	2	2	4
22	4	4	7	6	5	4
23	6	5	7	6	5	4
24	2	3	7	3	3	2
25	6	5	6	6	5	4
26	2	3	7	6	4	3
27	5	5	6	3	2	1
28	7	6	7	4	5	6
29	3	3	6	5	3	3
30	5	4	4	2	3	3
31	6	6	7	6	5	4
32	7	5	7	4	4	2

	Speech Rate Rating	Sys. Resp. Time Rating	Input Timeout Rating	Dbase Org. Rating	P/R Rating	Repeat Rating
1	7	7	5	5	4	•
2	4	3	7	4	1	•
3	4	7	6	7	•	•
4	2	3	5	1	1	•
5	4	5	6	5	4	6
6	5	4	5	6	•	5
7	2	5	7	7	•	•
8	5	2	4	6	1	6
9	5	7	5	6	•	5
10	5	6	6	3	4	4
11	2	4	7	7	1	1
12	3	4	4	5	•	•
13	5	7	7	7	•	7
14	6	4	7	6	•	4
15	4	4	5	7	•	•
16	1	1	7	6	1	1
17	4	3	6	7	1	•
18	5	2	6	6	1	•
19	6	4	6	6	1	2
20	6	6	6	6	•	•
21	7	7	5	5	•	•
22	5	6	6	7	•	•
23	4	3	7	6	•	6
24	7	7	6	5	•	•
25	6	3	4	5	1	3
26	2	3	6	7	3	•
27	6	6	5	5	1	1
28	3	6	3	6	•	1
29	6	5	4	6	1	•
30	4	4	3	3	1	•
31	6	3	7	7	•	4
32	5	2	6	7	•	1

	Spell Rating	Wallet Guide Rating	Music Rating
1	7	•	6
2	2	7	•
3	1	6	•
4	•	•	•
5	5	•	3
6	•	6	5
7	•	6	6
8	•	•	•
9	2	7	•
10	•	6	6
11	•	6	•
12	1	•	3
13	•	•	7
14	•	•	•
15	•	•	•
16	1	7	•
17	•	6	2
18	•	•	5
19	1	7	5
20	•	2	•
21	2	7	1
22	1	•	•
23	6	•	1
24	•	•	7
25	•	•	4
26	4	2	4
27	5	•	•
28	1	4	6
29	•	7	•
30	1	•	•
31	4	•	•
32	•	6	•

The vita has been removed from the scanned document