Assessing Human Performance Trade-Offs of a Telephone-Based Information System

by

Jimmy K. K. Wu

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

Walter W. Wierwille

Beverly H. Williges

August 4, 1989

Blacksburg, Virginia

ASSESSING HUMAN PERFORMANCE TRADE-OFFS OF A TELEPHONE-BASED INFORMATION SYSTEM

by

Jimmy Kwok-Kau Wu

Committee Chairman: Robert C. Williges
Industrial Engineering and Operations Research

(ABSTRACT)

Little research effort has been devoted to human interaction with telephone information systems. This study investigated the effects of system parameters and user characteristics on human behavior in an interactive telephone-based information system. The research method utilized a central-composite design to study four variables at five levels each. The four factors manipulated were: synthesized speech rate, time available for user input, subject age, and background music level. Subjects searched a fictitious department store database for 16 specific store items and transcribed 16 information messages which were spoken by a computer speech synthesizer. Subjective ratings of certain features of the system were solicited from the subjects and performance measures were also collected from the subjects on an on-line basis. Performance was evaluated by calculating regression equations relating the dependent measures and the independent variables. A response surface was plotted, and optimal settings for the information system were also calculated. Two seconds was found to be an optimal time for users to enter their selection. The computer synthesized speech rate should be set close to 120-150 words per minute. Background music or noise level should be kept below 50 dB(A); sound level above 50 dB(A) seriously affected user's ability to understand synthetic speech. Younger subjects (age 14 - 22) performed better in this study than older subjects (age 36- 62).

ACKNOWLEDGEMENTS

This research is supported by a grant form the National Science Foundation (contract # IRI-8604798) under the direction of Dr. Hal Bamford and Dr. Bruce Barnes. I would like to thank Dr. Robert C. Williges for his guidance throughout this project. I would also like to express my gratitude to Dr. Walter W. Wierwille and Beverly H. Williges, who provided insightful comments which greatly benefited this study. Dr. John G. Casali's help on the topic of noise measurement was also invaluable.

I would like to thank those who helped in this project:

, , and ; without their assistance, this study could not have been completed. Finally, I would like to thank my family for their unlimited support.

TABLE OF CONTENTS

INTRODUCTION 1
- Research Method 1
- Telephone Inquiry 2
- Experimental Overview 3
- Purpose 3
  - Age of User 4
  - Background music 4
  - Speech rate 4
  - Input time-out 5

LITERATURE REVIEW 6
- Interactive Telephone Information System 6
- Menu Retrieval 7
- Synthetic Speech Systems Guidelines 10
- Research Design 11
  - Second-order polynomial model 12
  - Speech rate 12
  - Age of User 14
  - Background music 15
  - Input time out 16

METHOD 17
- Experimental Design 17
  - Central-Composite Design 17
- Subjects 22
- Hardware 22
  - Computer 22
  - Speech synthesizer 24
  - Audiometer 24
  - Audio/Visual Instructions 25
  - Background Music. 25
  - Speaker telephone 26
- Database System 26
  - Keywords and target items 26
  - Information messages 28
- Objective Measures 33
  - Target search time 33
  - Total Number of Keypresses 34
  - Message transcription accuracy. 34
- Subjective Measures 34

RESULTS 36
Speech System 36
Search Time 36
Keypresses 51
Transcription Errors 56
Location of Transcription Errors 60
Optimal Settings 64
Subjective Ratings 64
Experimental Ratings 64
DISCUSSION 78
Input Time-out 78
Background Music Level 79
Speech Rate 81
Age 83
Input Time-Out by Music Level Interaction 83
Context Location of Information 84
Design Issues 85
Selection of Dependent Variables 85
Optimal Settings of Variables 91
CONCLUSIONS 93
REFERENCES 96
Appendix I 100
Appendix II 101
Appendix III 103
Appendix IV 105
Appendix V 106
Appendix VI 108
Vita 109

LIST OF TABLES

Table 1. Linear Transformations between Coded Values used in the Central-Composite Design and Real World Levels of the Four System Variables 21
Table 2. Experimental Conditions 23
Table 3. Syntax of the Information Messages 28
Table 4. Experimental Session 30
Table 5. Regression Equations for Subject Search Time 37
Table 6. Regression Equations for System Time 39
Table 7. Regression Equations for Total Search Time 40
Table 8. ANOVA Summary for Subject Search Time. 41
Table 9. ANOVA Summary for Total Search Time 42
Table 10. Regression Equations for Extra Keypresses. 53
Table 11. ANOVA Summary for Extra Keypresses. 54
Table 12. Regression Equations for Transcription Errors. 58
Table 13. ANOVA Summary for Transcription Errors. 59
Table 14. Optimal Settings of Independent Variables. 65
Table 15. Sums of Squares Summary for Subject Search Time 86
Table 16. Regression Equations for Subject Search Time using Raw Scores 88
Table 17. Regression Equations for Subject Search Time 89

LIST OF FIGURES

Figure 1. Example of a three-factor, second-order central-composite design....19
Figure 2. The Database hierarchy....27
Figure 3. Search Time vs. Age....44
Figure 4. Search Time vs. Input Time Out....45
Figure 5. Search Time vs. Music Level....47
Figure 6. Subject Search time vs. Speech Rate....48
Figure 7. System Time for the Input Time Out*Music Level Interaction....49
Figure 8. Subject Search Time for the Input Time Out*Music Level Interaction....50
Figure 9. Total Search Time for the Input Time Out*Music Level Interaction....52
Figure 10. Extra Keypresses vs. Age....55
Figure 11. Extra Keypresses vs. Music Level....57
Figure 12. Transcription Errors vs. Age....61
Figure 13. Transcription Errors vs. Music Level....62
Figure 14. Transcription Errors vs. Speech Rate....63
Figure 15. Difficulty in Understanding Information Messages....67
Figure 16. Background Music Level Ratings....68
Figure 17. Computer Speech Rate Ratings....69
Figure 18. Input Time-Out Ratings....71
Figure 19. Certainty Ratings in Transcribing Messages....72
Figure 20. Difficulty Ratings in Locating Target Items....73
Figure 21. Intelligibility Ratings....74
Figure 22. Naturalness of Computer Voice Ratings....75
Figure 23. Ease of Use Ratings....76
Figure 24. Menu Organization Ratings....77

INTRODUCTION

The advancement of computers and electronics has made speech technology an attractive option as a medium for human-machine communication. There are two main areas of interest in the arena of speech technology: speech recognition and synthetic speech production. Synthetic speech has achieved a relatively high degree of success in terms of intelligibility and in terms of its ease of production. Humans are much more adaptive in listening to speech produced by machines than are machines that try to interpret human speeches. For example, Greene et al. (1984) reported transcription error rates on Harvard meaningful sentences of synthetic speech using a DECtalk 1.8 system to be as low as 4.7 %. Even though the error rate is considerably higher than the natural speech transcription error of 0.8 %, nevertheless synthetic speech recognition by humans is promising. The purpose of this study was to investigate the effects of several variables on human behavior in an interactive telephone-based information system.

Research Method

This study was part of a research project which utilized sequential research methodology. The research project applied the sequential research design strategy in studying new and complex information systems. Most sequential designs are not the same in terms of statistical power as described by Williges (1981). There were three major stages in this particular sequential design. The first step was to select the critical variables that might produce significant effects on human performance. Second, one

must try to describe the relationships among the variables that were identified as critical. The last step was to seek an optimal performance level of the system by investigating different combinations of values of the variables.

In this particular project, the selection of the critical variables was conducted by five human factors engineers. A set of 95 system variables was initially selected to have possible effects on human performance in the telephone information system. However, in subsequent "brainstorming" sessions using expert and user opinions, literature reviews, and feasibility estimates; 16 variables were selected as having the greatest potential effects on human performance. Beaudet (1988) studied the 16 variables using a Hadamard (Diamond, 1982) design. The Hadamard matrix design allowed 16 variables to be examined using a total of 32 subjects. However, such a design allowed statistical analysis to be performed on main effects only. Therefore, the statistical methods used were efficient but not as powerful as most conventional human factors experiments. Beaudet's study served as a screening study in which the significant variables were subjected to more rigorous treatments.

This study utilized a central-composite design to investigate the relationship among four variables. Regression equations were calculated and performance measures were examined by examining the regression equations and by viewing the response surfaces. Optimal settings of the variables were also recommended.

Telephone Inquiry

The study focused on several behavioral issues relating to searching information in an auditory information system. Interactive telephone systems should involve two way communications between the user and the system. The information system should at least have the capability to output speech (either synthesized or recorded) and also allow for keypad input using the standard 12-key push button telephone. Communications are

then possible by a user pressing certain keys to effect certain responses from the system. Different types of interactive telephone information systems have been developed and evaluated. Some systems are available commercially such as IBM's Audio Distribution System as reported by Gould & Boise (1984).

Experimental Overview

Four variables in a telephone inquiry environment were manipulated in this study: age of the subject, background music level, synthetic speech rate, and time available for user input. Three of these variables were deemed to have affected user performance in the screening study conducted by Beaudet (1988) and were hence selected for further investigation in this study. The fourth variable, input time-out, was selected because Beaudet's dependent measures were not sensitive enough to detect any measurable changes of different levels of input time out. New dependent measures were added to re-evaluate the effects of input time-out. The levels and settings of each variable were calculated based on a four-factor, central-composite experimental design. The detailed experimental design is described in the method section.

Purpose

The purpose of this experiment was to describe the effects of four continuous independent variables on a number of dependent measures. Dependent measures were furthered divided into two categories: objective measure such as time required to complete a task, and subjective measure such as intelligibility ratings of the speech synthesizer. Regression equations of the four independent variables were calculated on the appropriate objective dependent measures. There were a number of advantages in studying the regression equations. First, significant main effects were examined by plotting the dependent measures against a selected range of independent variables.

Second, interactions between independent variables were examined by plotting response surfaces based on the regression equations. Response surface allowed performance trade-offs between two independent variables to be examined graphically. Third, optimal settings of the independent variables were recommended for designing the telephone inquiry system based on a number of measures.

Age of User

Subject age is an important factor in the evaluation of a telephone information system because older people have less experience in using computer systems than the younger population. Older people also exhibit longer reaction times in most motor tasks. Very little research had been conducted using age as a variable (Waterworth and Lo, 1984; Rosson and Mellen, 1985). Five different age group ranging from 13 to 62 were recruited. It was assumed that this age range would constitute a large part of the user population in interactive telephone information systems.

Background music

Background noise was manipulated in studies that involved perception of synthetic speech (Pisoni, 1979; Simpson and Marchionda -Frost, 1984). The effects of background noise or music on interactive tasks such as menu selection were poorly understood. In this study, five different levels of music ranging from 36.1 dB(A) to 65.4 dB (A) were manipulated in this study.

Speech rate

Several research experiments have reported different speech rate settings that the researchers claimed to yield the most intelligible speech. Simpson and Marchionda-Frost (1984) reported speech was most intelligible at a rate of 156 words per minute (wpm), while Merva (1987) reported 180 wpm seemed to be the best rate. Most of the

research experiments on synthetic speech manipulated only two or three discrete speech rate settings, and hence the optimal setting is simply the one on which most subjects performed best. With the use of the central-composite design, regression equations allowed speech rate to be studied at finer increments. Speech rates ranging from 120 to 240 wpm were used in this study.

Input time-out

Input time-out was defined as the time available for the subject to respond to the system and enter a keypress. A longer input time-out value allowed the user to have more time to decide on what action the user might take, but the entire task time would definitely be lengthened. In this study, input time-out was varied from two to ten seconds.

LITERATURE REVIEW

Interactive Telephone Information System

There are essentially two types of interactive telephone information systems; menu-driven systems and command driven systems. Command driven systems are often developed for specific applications such as voice-mail systems. Command driven systems usually require a long training period, and users are sometimes required to memorize the commands. Halstead-Nussloch (1989) developed a set of guidelines stating when a system should use command-based dialogue. He described four situations when a command-based dialogue are most appropriate:

- when command or function modes contains are few in number (five or less)
- each mode contains a large and complex structure of functions
- the user frequently needs to change modes within a session
- the phone-based interface is used frequently

Halstead-Nussloch (1989) concluded one should use menu-based dialogue in telephone information systems when there is a limited number of mode changes, when there is only a small number of commands, and when the system is used infrequently.

Menu-driven systems require less training time and are often used in tasks where most users are novices. Interactive information systems that employ the touch-tone telephone as a terminal is not a novel idea. A number of research projects have studied the feasibility of such systems. Witten and Madams (1977) developed a menu-driven telephone inquiry service which employed synthetic speech and allowed interactive information retrieval and data entry. Voice-mail systems are another example of interactive information services that are attracting a tremendous amount of

attention recently. Gould and Boise (1984) reported a complex voice-mail system that allowed the users to edit, send, and receive voice recorded messages to other users. Schmandt (1985a) developed a voice-mail prototype that accepts a wide variety of input inquiry and data-entry by defining each letter of the alphabet as a two-key combination of the 12-key standard telephone keypad.

One interesting fact about interactive telephone systems is that there is a great deal of user acceptance despite poor performance by the users (Wichansky, 1987). This sentiment is echoed by Schmandt (1985a). He reported that most users are tolerant of telephone-based information systems because they understand the limitations of such systems. Sperry computer (Anderson, 1984) tested a service that provides weather forecast service using synthetic speech. Users dialed into the system and entered the area code of the region for which they were interested in hearing the forecasts. User acceptance was reported to be as high as 97% for those who commented on the service.

Menu Retrieval

Numerous research projects have been conducted in the area of computer menu construction. Menu-driven software is a common way of communication between user and computer. Much research has been conducted with computer or videotext systems using visual menu interfaces. However, relatively little effort has been devoted to auditory database systems.

Halstead-Nussloch (1989) cited disadvantages of auditory menus that do not exist with visual database systems: auditory menu options must be spoken in a serial manner, the locations of menu options in the hierarchy are less obvious in an auditory system, and the user has to work with a limited keypad. One possibility of improving the quality of telephone-based interfaces is to improve the telephone itself to allow more

functions to be incorporated (e.g., add more keys to the keypad or a visual display to the unit). However, to modify the telephone is quite an impractical approach because most telephone-based information system are developed for a large user population, and the most efficient way is to make use of the standard touch-tone telephone.

Waterworth and Lo (1984) conducted an experiment using a keypad-input, voice-output automatic train time-table system designed for a standard 12-button telephone. The results were encouraging in that most subjects retrieved the correct information (train-times) in less than two minutes which was considered by the experimenter to be an excellent performance. Kidd (1982) conducted an experiment that prompted subjects to search for information in an auditory system. The auditory system was a direct translation of a videotext system (Prestel). She reported that most subjects had problems memorizing the database options, and most subjects were reported to have found the tasks to be difficult in one of the experiments. She reported that auditory database, when designed inappropriately, could placed a high burden of attention demand on the human short term memory and hence minimize the effectiveness of an auditory database.

Menu-driven systems often assume the user has little experience with the system and ask a series of questions prompting the user to perform simple keypress operations. For example, Waterworth and Lo (1984) tested a train time-table information system that queries or directs the user. If the user wants to go to the city of London or Ipswich, he or she simply presses either the "1" key for a positive response or the "0" key if the city spoken is not the user's chosen destination. It would be tedious and error-prone to enter the name of the city because of the increased number or keypresses required by the abbreviated keyboard. Inputting extensive textual (alphabetical) information via the telephone keypad is impractical because not all 26

letters of the alphabet are represented by the telephone buttons, and the exact same sequence of digits can represent more than one word. For example, names such as "Jim" and "Kim" are both represented by the same key sequence of "546" on a standard telephone keypad, and the letter "X" is not even represented by the 12 keys. Anderson (1984) reported that users had problems locating the "Y" and "N" keys on a keypad implying that searching for letters on the abbreviated keypad is not an easy task. Rosson and Mellen (1985) also tested a voice output (synthetic speech) system that provided traveling information within the city of Austin, Texas. They tested two types of keypad layout (spatial vs mnemonic) and concluded that subjects who were provided with the mnemonic layout performed better in locating the correct information by making fewer errors. High acceptance in using the service to locate entertainment and dining establishments was reported.

Engelbech and Roberts (1989) tested three different telephone-based interfaces. Subjects were to complete three tasks such as routing a call, screening a call, and retrieving a message using three different telephone systems. The three systems were: tone phone, prompt phone, and screen phone. A tone phone output responses in the form of tones which had different pitches, quality, and duration. The user entered commands via mnemonic codes on the telephone keypad, and received positive or negative tones in response. The prompt phone allowed users to move around the database using the keypad much like cursor keys, subjects received voice output for feedback and confirmation. Another version of the prompt phone displayed prompts, and users had to decide whether to select or skip the action prompted. The third system utilized a screen and mouse to complete the various tasks. Time and preference data both indicated that the screen phone was the best system. Though subjects that used the tone phone exhibited more errors in terms of usages, the tone phone was preferred over the prompt phone. This

implied that speed was considered by the subject as a very important factor because though subjects made fewer errors using the prompt phone; the long prompting messages were longer than the tones. As a consequence, the actual time it took to complete the tasks using the prompt phones was longer than the time it took using the tone phone.

The serial position of an item seems to have minimal effects on the accuracy of the selection as reported by Kidd (1982). In the study involving retrieval of information from a system composed of auditory menus, Kidd (1982) reported that the position of an item did not affect how the subjects behaved . She also stated there was no effect of list length on user response time; however, the longest list she tested consisted of only six items per list.

Efficiency of moving through auditory menus can be achieved by the appropriate design of prompting messages as reported by Aucella and Ehrlich (1986). In their study, user could quickly pre-empt a prompt by putting important contextual information in the beginning of a prompt. For example, the prompt "three messages are stored in your mailbox", is more concise than "there are three messages in your mailbox". An experienced user would immediately recognized that there are three messages in the mailbox after he or she had heard the first word of the sentence. Other ideas of speeding up dialogue flow included removing the entry of delimeters (similar to the function of return key on a keyboard), and also allowing the user to hang up the telephone at any time.

Synthetic Speech Systems Guidelines

Although most voice output database systems are different, they generally exhibit the same kind of problems. Issues such as option list lengths, training effects, and memory requirements of users are common interests of most researchers. There are

guidelines in existence (McCormick and Sanders, 1987; Schmandt, 1985b; Thomas *et al.*, 1984; Halstead-Nussloch, 1989) that address issues regarding the implementation of synthesized speech systems. However, most guidelines are extremely general and are quite tentative in nature. Most literature reported good user acceptance. However, very few reports on user performance and on user acceptance for real-life interactive systems exist.

Research Design

This study manipulated four independent variables with five levels each. The specific design is called a four factor, orthogonal central-composite design. Central-composite designs were originally developed to determine the optimal combination of various factors in chemical processes in attempts to maximize yield (Box and Wilson, 1951). Williges and his colleagues have applied central-composite designs in conducting various human factors experiments (Williges, 1981). The advantage of central-composite design is that a quantitative functional relationship such as a response surface between human performance and system parameters can be established and evaluated. Other experimental designs often allow the researcher to interpret the results only as significant or non-significant between different levels of the variables.

Williges and North (1973) used response surfaces to evaluate the performance of human operators locating designated target symbols on a series of maps displayed on black-and-white and color television monitors. Williges and Williges (1982) used a similar design to study 22 separate dependent variables in a data entry task. In the data entry study, three classes of metrics (operator satisfaction ratings, work-sampling procedures, and embedded performance measurement) were deemed to be important measures in evaluating human-computer interfaces of data entry tasks. The three types of metrics yielded a total of 22 dependent variables. Four independent variables were

manipulated, and subsequent data analysis using response surface methods indicated system delay time and keyboard echo rate were two variables that affected user performance the most.

Second-order polynomial model

Williges (1982) suggested that second order polynomials are useful in assessing user performance with computer-based systems. The polynomial equation predicting human performances is in the form:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \sum_{i=1}^{k} \beta_{ii} X_i^2 + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \beta_i \beta_j X_i X_j + \varepsilon$$

where Y is some dependent measure of human behavior. It is expressed in terms of the intercept value β_0 and the linear combinations of various system parameters (X's). The number of system variables manipulated is represented by k. ε in the equation represents the estimated error in prediction. The Beta (β) weights in the equation give a clear indication of how much variance a particular component has contributed towards the total variance if the design is orthogonal. Therefore, important system parameters (X's) can be readily observed.

Speech rate

Simpson and Marchionda-Frost (1984) reported that subjects preferred a speech rate of 156 words per minute (wpm) as the most intelligible speech rate by a speech synthesizer. In their experiment, a video game was used to simulate flight tasks while the subjects had to report threat messages. The threat messages were produced by a Votrax ML-1 synthesizer. Performance measures such as mean response time to report threat messages showed no differences for the three speech rates used (123,

156, 178 wpm). However, subjective ratings reflected that most users preferred a speech rate of 156 wpm over either the slower or the faster rate.

Slowiaczek and Nusbaum (1985) reported a significant transcription accuracy decrement when speech rate was increased from 150 wpm to 250 wpm. They used a Speech Plus Prose 2000 synthesizer and discovered that percentage correct identification for meaningful and anomalous sentences dropped drastically. The percentage of correct transcriptions averaged around 90% for the 150 wpm rate as compared to about 60% for the 250 wpm speech rate. However, their recommendation of 150 wpm as an optimal speech rate must be taken with caution. In this experiment, there is clearly evidence to support the idea that a speech rate of 150 wpm is more intelligible than a speech rate of 250 wpm. However, some intermediate speech rate might produce better performance within the 150 to 250 wpm range.

Waterworth and Lo (1984) evaluated six different speech rates (63, 82, 103, 121, 130, 150 wpm) and reported that all six different speech rates were equally intelligible using a custom built synthesizer (VB-1 by British Telecom). However, there is no evidence from this study to suggest that speech rates higher than 150 wpm would produced sub-optimal performance.

In an attempt to investigate how human perceived fast speech rate, Herlong (1988), using a DECtalk 2.0 speech synthesizer, conducted an experiment asking subjects to retrieve information about different items in an fictitious department store database. Speech rates of 180 wpm and 240 wpm were used. The mean transcription error rate for 240 wpm was 10.81%, as compared to 6.63% for the 180 wpm condition. Merva (1987) tested transcription accuracy on speech rate and repetition of messages spoken by synthesized speech using the same speech synthesizer as in the Herlong (1988) study. In Merva's study, three speech rates were used: 150, 180, and

210 wpm. With one or two repetitions of the messages, error rates were 12 and 3 percent respectively for the 180 wpm test conditions. However, the speech rates of 150 and 180 wpm were not statistically different in terms of transcription errors but error rates of 150 and 180 wpm were significantly lowered than that of 210 wpm. According to the experiments described above, an optimal setting of speech rate seems to lie above 150 wpm but below 210 wpm. Finer increments of speech rate settings are required to pinpoint a speech rate that would produce the highest intelligibility. Interestingly, the rate of 180 wpm closely approximates the rate of human speech (Lee, 1983).

Age of User

Age of the user is an important factor in designing an interactive database system. First, older subjects might have hearing loss that hinders their listening ability. Second, limited exposure to computers in the older population might pose special problems such as understanding the hierarchial nature of many database systems. Little research on synthetic speech systems has been devoted to older subjects.

Waterworth and Lo (1984) used a simulated train time-table information system to evaluate the effects of age. Four different synthesizers (Microspeech-2, Prose-2000, Votrax CDS-II, British Telecom VB-1) and one natural voice were used. Five age groups (under 20, 20-29, 30-39, 40-49, 50-64) were tested in two separate experiments and no performance differences between the five groups were found. However, in another attempt to study how different age groups behaved in using speech systems, Rosson and Mellen (1985) reported that older subjects took longer to recover from error. For example, older persons were more likely to use conservative strategies, repeating the same unsuccessful actions before taking a new approach. However, in their study, only 18 subjects were used with ages ranging from 18 to 38 years old. There was no report of how the subjects were bracketed into different age-

groups, and the oldest subject was only 38. Therefore, the reported effects of age might be an artifact of individual performance differences. Beaudet (1988) studied college age subjects (18-30) and an older age group (45-60). He found differences in performance such as time to retrieve information and number of erroneous keypresses. In this study, the effects of age were studied using five different age groups.

Background music

Noise induced in the communication channel or in the background are known to produce degradation of speech intelligibility (Kryter, 1972). Little research effort has been devoted to studying on the effect of noise on human performance of interactive database systems using synthetic speech.

Pisoni (1979) reported background masking noise might contribute to poor performance in tasks that involve phoneme recognition. Simpson and Marchionda-Frost (1984) tested pilot performance by presenting synthetic threat warning messages in a cockpit environment. White noise was introduced purposely to simulate the noise in an helicopter cockpit environment. They hypothesized that with the presence of noise in a channel, communication quality would increase if higher pitch signal is selected. The results of the experiment proved that subjects exhibited no preference in voice pitch levels in the presence of noise. However, no data were reported on how noise had affected performance.

Little research effort has been devoted to time-varying noise in either the background or the communication channel. Recommendations of maximum allowable background noise for various activities (job-shop, offices) for design purposes are well documented (Beranek et al., 1971). Whether these guidelines will apply to the use of interactive databases using synthetic speech remains unknown.

Beaudet (1988) used background music during half of the experimental sessions involving information retrieval in telephone information systems. He reported that background music produced effects on user performance such as time to retrieve relevant information and the number of erroneous keypresses resulted.

Five different sound levels of music were used in this study to investigate the relationship between level of background music and user performance. Time-varying noise was a variable of interest since noise in a real-world setting for the telephone inquiry system would most likely have that characteristic.

Input time out

Input time out is the time available for the subject to invoke a keypress before the system presents another menu choice. Beaudet (1988) reported that input time-out values of 2 and 4 seconds produced no noticeable effects on performance as measured by the number of keying errors, search time ratio, and search efficiency ratio. However, he failed to analyze the average search time to complete a search. In his study, objective measures were recorded as search time ratio comparing the time taken to retrieve a piece of information to the time of a simulated expert who made no errors. The decision to exclude average search time to complete target searches was based on the assumption that a system providing longer input time-out periods would lead to longer total search time; hence the ratio scoring schemes were used to evaluate user contributions to search time. However, average time to complete a search is a relevant measure because using this measure, one can evaluate how different settings of input time-out would affect the actual amount of time the user needed to complete the searches as a result of both system characteristics and user strategies.

METHOD

This study required subjects to retrieve and transcribe information in a fictitious department store database. Database items to be located by the subject were presented on a computer terminal. The subject used a touch-tone telephone and synthetic speech to interact with the database. The task included a series of target searches in which the subject was required to locate a particular item in the department store database system. Once the item was located, the subject was required to transcribe an information message associated with the item. The idea of the information message was to mimic a system in which users could obtained information such as prices, locations, and availability of certain store items.

Experimental Design

The experimental design proposed is an orthogonal, four-factor, between-subjects, central-composite design. Four variables were studied: synthetic speech rate, age of the subject, background music level, and length of the input time-out period.

Central-Composite Design

Central-composite designs allowed regression equations of dependent variables (in most cases, human performance measures) to be calculated. For example, a regression equation for error rate could be calculated in terms of all first and second order components of the four independent variables. Central-composite designs are particularly suitable for experiments when third and higher order effects are assumed to be negligible. To view performance trade offs, transect plots or response surfaces

could be achieved by plotting one dependent measure on the ordinate axis against two independent variables on the horizontal axis.

Figure 1 is an example of a three-factor, second-order central-composite design to evaluate automobile driving performance (Y) as a function of wind gust characteristics of X_1, X_2, X_3 (Williges, 1981). Note that the design is composed of a full factorial portion (eight data points in this case), six additional data-points as probes, and one center point. Therefore, one can readily observed that five levels of each factor are needed. There are two levels for the factorial portion (-1, +1), two levels for each of the probe portion ($-\alpha$, $+\alpha$), and a center point with levels set at (0,0,0). The five levels are ($-\alpha$, -1, 0, +1, $+\alpha$). The calculation of the α is quite crucial and will be explained below.

Williges (1981) described that in a second-order, central-composite design; the number of data points is calculated by

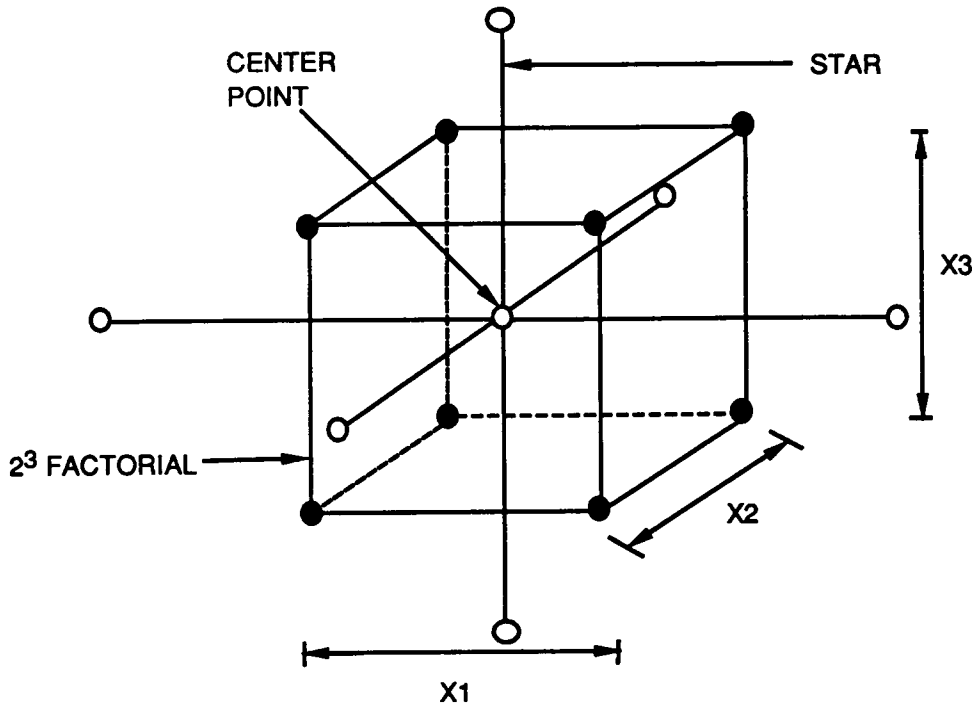
$$T = F + 2K + C \qquad (1)$$

where F = number of data points of the factorial portion of the design. In this study, there were four factors with each factor varied at two distinct level in the factorial portion of the experiment. Therefore, the F value is $4^2 = 16$. k is the number of factors in the design (4 in this study). C is the number of replications at the center point which equalled 1 in this study. From Equation 1:

$$T = 16 + 2(4) + 1 = 25$$

Therefore, the number of unique data points is 25. In a four factor central-composite design, each factor is studied at 5 levels ($-\alpha$, -1, 0, +1, $+\alpha$) where the value of α depends on each specific design. β weights in an orthogonal design do not interact with

DESIGN CONFIGURATION



Treatment Combination	X_1 Wind Gust Frequency	X_2 Wind Gust Velocity	X_3 Wind Gust Direction
1	+1	+1	+1
2	+1	-1	+1
3	+1	+1	-1
4	+1	-1	-1
5	-1	+1	+1
6	-1	-1	+1
7	-1	+1	-1
8	-1	-1	-1
9	$+\alpha$	0	0
10	$-\alpha$	0	0
11	0	$+\alpha$	0
12	0	$-\alpha$	0
13	0	0	$+\alpha$
14	0	0	$-\alpha$
15	0	0	0

Figure 1. Example of a three-factor, second-order central-composite design to evaluate automobile driving performance (Y) as a function of wind gust characteristics (X_1, X_2, X_3) (Adapted from Williges, 1981; page 67)

each other; that is, the correlation between β weights must be 0. In an orthogonal design α is calculated by:

$$\alpha=(\frac{QF}{4})^{1/4} \qquad (2)$$

where $Q=[(F + 2K + C)^{1/2} - F^{1/2}]^2$ (3)

In this design, $Q=[(16+2(4)+1)^{1/2} - 16^{1/2}]^2 = 1$. The α value of the design is calculated by $\alpha=(\frac{QF}{4})^{1/4} \Rightarrow \alpha=(\frac{1(16)}{4})^{1/4} = \sqrt{2}$. The prescribed α value is then $\sqrt{2}$, calculated for a four factor, orthogonal design with one center point.

Experimental settings were calculated by performing linear transformations using the coded values ($-\sqrt{2}$, -1, 0, +1, $\sqrt{2}$). In this study, it was determined that speech rate was to be studied between 120 words per minute (wpm) and 240 wpm. The 120 wpm condition corresponded to the $-\sqrt{2}$ coded level and 240 wpm represented the $+\sqrt{2}$ coded level. The 0 coded level was the midpoint between the two extreme values, hence it was set at 180 wpm. Since 240 wpm is a $\sqrt{2}$ unit increase from the center point of 180 wpm, a one unit increase corresponded to a setting of 222 wpm. Similarly, a one unit decrease from 180 wpm yielded a setting of 138 wpm. The five real-world speech rate settings were then calculated to be: 120, 138, 180, 222, and 240 wpm.

See Table 1 for a listing of the coded values and the transformed real-world values of the four independent variables to be used.

Efficiency of data collection was obtained by sacrificing three way or higher interactions. In this experiment, four variables were studied at five different levels using only 25 data points, whereas a full factorial between-subject design would

Table 1. Linear Transformations between Coded Values used in the Central Composite Design and Real-World Levels of the Four System Variables

	Levels of the Four Independent Variables				
	-1.414	- 1	0	+1	+1.414
Speech Rate (wpm)	120	138	180	222	240
Input Time-Out (seconds)	2	3	6	9	10
Background Music (dB(A))	36.1	40.5	50.8	61.1	65.4
Age (years)	13-17	20- 24	36 -40	52 - 56	58 - 62

required 625 different data points. The 25 experimental conditions are presented in Table 2.

A between-subject design was selected because one of the factors that was studied was age, and hence a within subject design was not possible.

Replication across the entire experiment was deemed appropriate to increase the power of the statistical analysis, specifically, increasing the degrees of freedom of the error term. Therefore, a total of 50 subjects were required for this experiment.

Subjects

Fifty subjects divided into five different age groups were needed in this study. Subjects were recruited on a volunteer basis and were compensated for their participation. It was expected that most people that use telephone information system would be very likely be between the age of 15 to 60. Subjects from age 15 to 60 were recruited for two additional reasons: first, subjects that are older than 60 years old may possess some kind of hearing impairment; second, subjects that are younger than 15 years of age may not understand the keywords of the database system. However, it was often difficult to obtain subjects at the exact age prescribed by the values of the central-composite design. Therefore the age criterion was relaxed to plus or minus two years of the prescribed age. The five age groups were 13 to 17, 20 to 24, 36 to 40, 52 to 56, and 58 to 62 years old. All subjects were native English speakers who had little exposure to synthetic speech systems.

Hardware

Computer

A Vax 11/750 mini-computer manufactured by Digital Equipment Corporation (DEC) was used for the entire study. Information was displayed using a DEC VT-220

Table 2. Experimental Conditions

	speech rate (wpm)	input time-out (sec)	background music (dB (A))	age (years)
1	222	9	61.1	54
2	222	9	61.1	22
3	222	9	40.5	54
4	222	9	40.5	22
5	222	3	61.1	54
6	222	3	61.1	22
7	222	3	40.5	54
8	222	3	40.5	22
9	138	9	61.1	54
10	138	9	61.1	22
11	138	9	40.5	54
12	138	9	40.5	22
13	138	3	61.1	54
14	138	3	61.1	22
15	138	3	40.5	54
16	138	3	40.5	22
17	240	6	50.8	38
18	120	6	50.8	38
19	180	10	50.8	38
20	180	2	50.8	38
21	180	6	65.4	38
22	180	6	36.1	38
23	180	6	50.8	60
24	180	6	50.8	15
25	180	6	50.8	38

terminal. This experiment had real-time priority over any other activities of the computer. The mini-computer also recorded and time-stamped all transactions made between the user and the system.

Speech synthesizer

DECtalk (version 2.0) was the speech synthesizer used in this study. It is a rule-based text to speech system manufactured by Digital Equipment Corporation. The voice that was used in this study was Perfect Paul with all the parameters at default settings. DECtalk's Perfect Paul voice was selected because of its high intelligibility. Words or sentences that were used in this study were not modified using any coding schemes such as phoneme entry. There were no attempts to manipulate parameters such as pitch or any other voice qualities to make the words or sentences sound more realistic. Words and sentences were typed into the DECtalk system with no manipulation; the only exception was that a hyphen was inserted in compound words such as basketball in order to avoid improper pronunciation that might have resulted.

Audiometer

A Beltone 109 Audiometer was used to screen subjects for any serious hearing loss. A hearing test was administered to the subject with the following provisions for passing: subjects must have heard interrupted tones presented at 26 dB(A) and at frequencies of 750, 1000, 2000, and 4000 Hz. The criterion was relaxed at the 4000 Hz level, where subjects that had failed to hear the tones at 26 dB(A) were given additional tones at 30, 35, and 40 dB(A). Only subjects that heard the tones at or below 40 dB(A) were allowed to participate. The sound level at which the subjects heard the tones was recorded. The purpose of the hearing test was to screen out subjects with poor hearing. The hearing test criteria was relaxed at the 4000 Hz level because people's

hearing ability tend to deteriorate with age. Spoor (1973) using previously gathered data from five different studies with a total sample size of 7724, calculated that for males between the age of 50 - 59, a quarter of the male population have a hearing threshold of 33 dB at 4000 Hz. Maurer and Rupp (1979) reported hearing threshold for males of age 55 - 64 at 4000 Hz to be slightly above 45 dB. Their data were gathered from an extensive National Health survey with a sample size of 6000 people. There appeared to be a sharp decrease of hearing threshold at frequencies greater than 2000 Hz. Men possessed hearing loss much more severe than women. The female hearing threshold at 4000 Hz was slightly above 20 dB (Maurer and Rupp, 1979).

Audio/Visual Instructions

Instructions in this study were presented visually using pre-recorded videotapes. A GE stereo video cassette recorder and a stereo television set were used to present the instruction tapes. A JVC GX-5700 video camera was used to monitor the entire experiment.

Background Music.

A Realistic STA-19 stereo receiver, a Realistic SCT-45 stereo dubbing cassette deck, and a pair of Realistic Minimus 7 speakers was connected to furnish the background music. Background music was provided by playing a 45-minute audio cassette tape containing seven different songs recorded by the Muzak company. The volume control dial of the stereo receiver consists of 40 discrete detents. A three minute segment of the 45-minute taped was integrated using a Larson-Davis Precision Integrating Sound Level Meter (Model 800B). The first 13 detents were integrated using the equivalent noise level(L_{eq}) method as described in Williams (1978). The entire tape (45 minutes) was then measured again at the ninth and the thirteenth

detents. It was discovered that the music level of the entire tape (45 minutes) was 1.5 dB(A) higher than the three minute integration of all 13 detent levels. Therefore, 1.5 dB(A) was added to each value obtained from the three minute integration. The five detents that produced sound levels closest to the prescribed sound levels of 36.1, 40.5, 50.8, 61.1, 65.4 dB(A) were 0, 3, 6, 11,12 clicks respectively. Sound level was then set at each experiment session by selecting the appropriate detent number.

Speaker telephone

A touch-tone speaker phone was used to present the output of the information system. The speaker telephone was a Panasonic Easaphone (VA-8205). The telephone volume was not altered and was set at approximately 64 dB(A), which was measured near subject's ear height.

Database System

The database system was composed of three major components: target items, keywords, and information messages. Each subject was prompted to search for 16 target items such as "chicken cookbooks", "golf books", and "women's cotton blouses". The system then spoke a menu of keywords which were composed of items such as furniture, appliances, music, etc. The database system was constructed with 8 menu items on each of 2 levels (8x2). Therefore, the database consisted of 64 target items (store items).

Keywords and target items

Figure 2 depicts the database hierarchy. A considerable amount of effort was expended in minimizing semantic ambiguity in the database structure (Beaudet, 1988). Preliminary experiments were carried out to ensure that each target item would be found under the applicable keywords. For example, target item "blazers" is related to

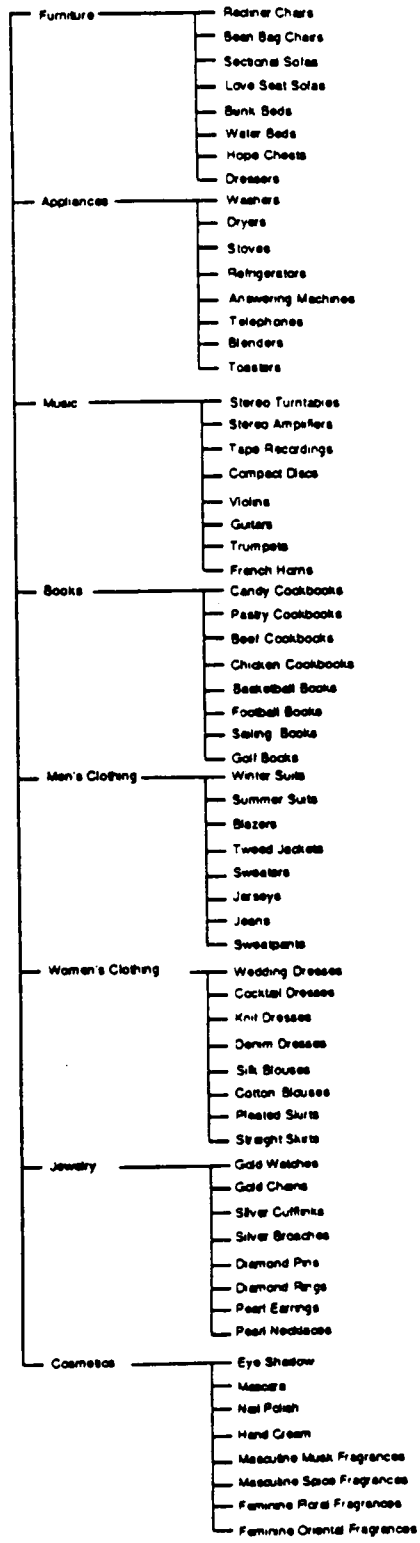


Figure 2. The Database hierarchy

the keyword "men's clothing". Vague or unintelligible keywords were replaced prior to use in any study.

Information messages.

One unique information message was associated with each target item. The purpose of the information message was to measure how well people understood synthetic speech. Each information message conformed to a fixed format of "*Modifier Subject Verb Preposition Modifier Object*". For example, the sentence "coaches biographies are reduced for quick clearance" is an example of the information message that conforms to the above rule. The subject was required to transcribe the information message by typing the messages using the computer terminal's keyboard. Only the modifier, subject, modifier, and object were scored. Each information message had its own unique modifiers, subject, and object. Therefore, no two sentences tested were alike in terms of scored words. However, verbs and prepositions did repeat in the middle part of the sentence and hence they were not scored. The messages provided four types of information: location, price, availability, or information. The sentence structure is presented in detail in Table 3. See Appendix I for the actual information messages associated with each target (store) item.

Procedure

Table 4 is a brief summary of the experimental procedures that was used in this study. The experimental session was composed of four major phases: welcome and orientation, instructions and practice, experimental task, and post experimental debriefing.

Table 3. Syntax of the Information Messages

Information Type	Syntax
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

Table 4. Experimental Session

1.) WELCOME AND ORIENTATION (approximately 15 minutes.)

Obtain Informed Consent
Administer Subject Information Questionnaire
Conduct Hearing Test

2.) INSTRUCTIONS AND PRACTICE (approximately 20 minutes.)

Present Audio-Written Introductions
Present Audio-Written Instructions
Present Video Instructions
Present Keypad Instructions
Explain Subject Tasks
Practice on Two Target Searches

3.) EXPERIMENTAL TASK (approximately 30 minutes.)

8 Experimental Targets

Target Search
Transcription of Information Messages
Target Ratings

Break (mandatory 1 minute)

8 Experimental Targets

Target Search
Transcription of Information Messages
Target Ratings

Post Experimental Ratings

4.) POST EXPERIMENTAL SESSION (approximately 15 minutes.)

Debriefing
Payment and Dismissal

An informed consent document was given to each subject detailing the purpose of the experiment and the subject's rights. The informed consent document is presented in Appendix II. An information questionnaire was provided to record the demographics of the subject. Information such as age, hearing, and regions of rearing was recorded by subject's self report.

An introduction was presented to the subject. The introduction described the tasks to the subjects in a very general way, followed by step-by-step instructions that detailed the entire experiment. The instructions were presented orally by the DECtalk, and type-written instructions were also provided. The purposes of the audio instructions were two-fold; to familiarize the subject with the synthetic voice and also to reduce the possibility that the subject intentionally skip portions of the instruction. The subject also watched a videotape that portrayed the two practice target searches. Audio instructions (with written instruction as supplement) were presented to the user in order to relate the functions of each telephone key. After the presentation of the keypad instructions, the experimenter answered any questions by the subject and the experimenter asked the subject to recap the steps of the task. The subject's summary was essential in order to avoid any misinterpretations of the instructions. See Appendix III for the complete audio instructions.

Each subject was required to search for a total of 16 targets. Target items were store items that were in the database system. Some examples of target items were washer, compact disc recordings, silk blouses, etc. In the beginning of each search, the computer terminal displayed the message "What is the information message for < *target item* >?". About fifteen seconds after the message was displayed, the target message would disappear and a prompt "ready..." was displayed. Two seconds after the "ready..." prompt, a "begin the search for the store item" message was displayed on the screen.

This was a cue to notify the subject that the phone system was about to begin speaking the keywords.

There were eight keywords in the top level menu of the database; each keyword contained an additional eight store items in its second menu level. The subjects were instructed to use the number key (#) to select the keyword just presented. The subjects had the option of going back one level in the dialogue (e.g., from store items back to keyword) by using the asterisk (*) key. Subjects also had the ability to restart from the beginning of the hierarchy by pressing the zero (0) key.

Once the subject had reached a store item, there were two possible scenarios. First, the user might have found the incorrect store item; in such a case, a message such as "at store item eye shadow, continue searching" would prompt the subject that eye shadow was not the right choice. If the subject had located the correct store item, the subject was instructed to press the "2" key on the telephone keypad to listen to the information message. The message described price, location, availability or information about shopping in the department store. The message content was not necessarily directly related to the specific store item; for example, if the subject had located women's cotton blouses, the information message might be "local addresses are required on bank checks". Once the message was spoken, the computer prompted the user to transcribe the message by saying "Begin Transcription". Subjects had an unlimited amount of time to type their answer via the keyboard of the terminal.

After each target search, subjects were instructed to rate three quality measures of the target search. The three ratings are: transcription certainty, transcription difficulty, and store item search difficulty. The same procedure was the same for all sixteen targets with a mandatory one-minute break inserted after eight target searches.

When all 16 target searches had been completed, the subjects rated seven characteristics of the telephone system such as, ease of use of the information system, intelligibility of the computer voice, naturalness of the computer voice, speech rate of the computer voice, time available for user input (input time out), the effect of background music, and menu organization of the database. All post-experiment ratings used a similar 7 point bipolar scales. The subject debriefing was conducted, and questions from the users were answered. The subjects were then thanked for their participation and paid upon the completion of the debriefing session.

Objective Measures

Three objective measures were used in this study along with ten subjective rating measures. The objective measures are explained in detail below.

Target search time

Success of one target search can be expressed as the total time a subject spent searching for each of the 16 targets. It would be erroneous just to compare the total search time between the different experimental conditions because different subjects were assigned different input-time out values. For example, a subject that had been assigned an input time-out value of 10 seconds would require more time to perform the task in a much longer time than a subject that had been assigned a value of two seconds even if their performances were similar. This is true because menu items were presented serially and if an item spoken were not the desired item, a subject had to wait a period of time equal to the length of the pre-set input time-out before the next item was spoken. Therefore, a short input time-out allowed a subject to finish the search in a shorter time.

To compare search times between conditions meaningfully, the system's contribution to user search time must be extracted. System contribution (time) was calculated by running four iterations of a simulated search. The simulated search time composed of two distinct timing parameters: time available for user input (input time-out), and output time of the speech synthesizer (i.e., the time necessary for the synthesizer to say a word or sentence). Because the operating system of the computer (VAX/VMS) and the speech synthesizer both exhibit small variances in the amount of time to respond or to speak the keywords, four iterations of the simulation were averaged to produce a better estimate of system contribution to search time. A subject's contribution to search time could be compared by subtracting the portion of system time from total search time.

Total Number of Keypresses

The minimum number of keypresses required to complete a search in this experiment is three. The subject had to press at least two keys to reach the desired item, and the subject had to press one key to hear the information message. Any extra keypresses made by the subject could be deemed as a valid measure of errors.

Message transcription accuracy.

As described earlier, the first and last two words in each information message transcription were scored. Since the middle words in all information messages were redundant prepositions and verbs, they were not included in the scoring scheme.

Subjective Measures

Subjects were instructed to rate three quality measures following each of the 16 target searches as described previously. The ratings were recorded using a seven-point verbally anchored scale. First, subjects were prompted to rate how certain they were of

their transcription. Second, subjects were prompted to rate how difficult it was to understand the information message, and the last rating registered the difficulty of searching for the store item. The rating scales for these three measures are presented in Appendix IV.

Characteristics of the telephone system were assessed by asking subjects to complete six additional subjective ratings at the end of each experimental session:

- ease of use of the information system
- intelligibility of the computer voice
- naturalness of the computer voice
- speech rate of the computer voice
- time available for user input (input time out)
- effect of background music
- menu organization of the database system

The ease of use rating was solicited from the subject in order to obtain an overall assessment measure that would evaluate the acceptance level of the telephone based information system. The only characteristic of the synthetic speech that was manipulated in this experiment was the rate of presentation. The speech rate factor might produce noticeable effects on human perception of the intelligibility and naturalness of the system. Subjects were asked to rate the variables that were manipulated explicitly in this study. Ratings of speech rate, input time-out, background music ratings were also requested from the subjects. All post-experiment ratings were solicited using seven point, verbally anchored bipolar scales. Appendix V presents a listing of the post-experiment ratings and the descriptive anchors.

RESULTS

Fifty-two subjects participated in this study. Fifty-one of the 52 subjects passed the hearing test; though eight of the 52 subjects had hearing threshold greater than 26 dB at a frequency of 4000 Hz, all subjects were able to hear the test tones of 4000 Hz at a sound level of 40 dB or lower. One subject failed the hearing test at the frequency of 4000 Hz even when the sound level was raised to at 40 dB. The data of the subject were discarded and were not included in the analysis. Another subject's data were not included because the subject had severe problems with spelling during the transcription tasks. This subject had heard and understood most of the sentences but was unable to transcribe the sentences accurately because of spelling errors. The data from the remaining 50 subjects were retained and were used in this study.

Regression equations for each of the three types of dependent variables were calculated. Significant linear or quadratic predictors are presented by plots depicting the effects predicted by the regression equations. Significant interaction effects are illustrated by response surface plots. Optimal settings of the independent variables are also discussed in this section.

Speech System

Search Time

The time it took a subject to retrieve an item can be broken down into two distinct portions: subject search time and system time. System time was obtained by computer simulation as described in the method section. Subject search time is then calculated by subtracting system time from the total time obtained from each subject's

data. Total search times varied from a minimum value of 0.54 minute/trial to an maximum value of 2.69 minute/trial. The average total search time was 1.06 minute/trial (σ^2=0.42). On the average, 0.86 (σ^2=0.31) minute of each search was contributed by the system. Subject search times ranged from 0.02 to 1.52 minute/trial (σ^2=0.25).

Regression equations for subject search time, system time, and total search time are presented in Table 5-7. β weights can be readily observed by examining the regression summary tables. Table 8 and Table 9 are Analysis of Variance (ANOVA) summary tables for subject search times and total search times, respectively.

The regression model has a R^2 value of 0.4626 for subject search time. From Table 5, four variables were shown to be significant. The error term used was calculated by separating the subject within-treatment variance from the total variance and has 25 degrees of freedom, the remaining 10 degrees of freedom were partitioned into the Lack-of-Fit term. Lack-of-Fit was not significant for subject search time ($F_{10,25}$=0.514, $p>0.1$). Therefore, it was valid to assume that third or higher order effects were negligible. Four variables were found to be significant predictors: speech rate ($F_{1,25}$=5.665, $p<0.05$), music level ($F_{1,25}$=4.350, $p<0.05$), age ($F_{1,25}$=5.294, $p<0.05$), and the interaction of input time-out by music level ($F_{1,25}$=4.793, $p<0.05$).

The variable system time was predicted almost entirely by the variable input time-out (R^2=0.9995). The only other variable that could have affected the simulated system time was speech rate. An analysis of variance was conducted on the two factors, and 99 percent (4.7913/4.8399) of the variance was accounted for by the variance of the input time-out factor .

Table 5. Regression Equations for Subject Search Time

Variable	df	Estimate (β)	Std. Error	T ratio	Prob
Intercept	1	0.133	0.092	1.444	0.1577
Speech Rate (S)	1	0.088	0.034	2.565	0.0148
Input Time-Out (I)	1	0.028	0.034	0.806	0.4255
Music Level (M)	1	0.077	0.034	2.248	0.0310
Age (A)	1	0.085	0.034	2.479	0.0181
S^2	1	0.010	0.054	0.186	0.8534
I^2	1	0.002	0.054	0.031	0.9756
M^2	1	0.004	0.054	0.080	0.9366
A^2	1	0.067	0.054	1.236	0.2246
SI	1	0.029	0.038	0.753	0.4565
SM	1	0.045	0.038	1.169	0.2503
SA	1	0.053	0.038	1.377	0.1774
IM	1	0.091	0.038	2.359	0.0240
IA	1	0.021	0.038	0.555	0.5822
MA	1	0.025	0.038	0.655	0.5170

Table 6. Regression Equations for System Time

Variable	df	Estimate (β)	Std. Error	T ratio	Prob
Intercept	1	0.851	0.004	230.612	0.0001
Speech Rate (S)	1	-0.035	0.001	-25.167	0.0001
Input Time-Out (I)	1	0.346	0.001	251.600	0.0001
Music Level (M)	1	0.000	0.001	-0.038	0.9696
Age (A)	1	0.000	0.001	-0.035	0.9722
S^2	1	0.007	0.002	3.029	0.0046
I^2	1	0.000	0.002	0.102	0.9193
M^2	1	0.000	0.002	0.147	0.8675
A^2	1	0.000	0.002	0.168	0.9562
SI	1	0.000	0.001	-0.055	0.9562
SM	1	0.000	0.001	-0.001	0.9993
SA	1	0.000	0.001	0.017	0.9862
IM	1	0.000	0.001	0.034	0.9731
IA	1	0.000	0.001	0.004	0.9968
MA	1	0.000	0.001	-0.010	0.9924

Table 7. Regression Equations for Total Search Time

Variable	df	Estimate (β)	Std. Error	T ratio	Prob
Intercept	1	0.984	0.093	10.552	0.0001
Speech Rate(S)	1	0.054	0.035	1.543	0.1318
Input Time-Out(I)	1	0.374	0.035	10.752	0.0001
Music Level (M)	1	0.077	0.035	2.223	0.0328
Age (A)	1	0.085	0.035	2.453	0.0193
S^2	1	0.017	0.055	0.304	0.7629
I^2	1	0.002	0.055	0.034	0.9727
M^2	1	0.005	0.055	0.085	0.9326
A^2	1	0.068	0.055	1.230	0.2268
SI	1	0.029	0.039	0.743	0.4624
SM	1	0.045	0.039	1.157	0.2550
SA	1	0.053	0.039	1.363	0.1815
IM	1	0.091	0.039	2.337	0.0253
IA	1	0.021	0.039	0.550	0.5959
MA	1	0.025	0.039	0.647	0.5215

Table 8. ANOVA Summary for Subject Search Time

SOURCE	DF	SS	MS	F	p ≤
Regression	(14)	1.42673438	0.10190960	1.853	
Speech (S)	1	0.31151470	0.31151470	5.665	0.05
Input time out (I)	1	0.03078852	0.03078852	0.560	
Music Level (M)	1	0.23921032	0.23921032	4.350	0.05
Age (A)	1	0.29109811	0.29109811	5.294	0.05
S*S	1	0.001640069	0.001640069	0.030	
I*I	1	0.00004478	0.00004478	0.001	
M*M	1	0.00030442	0.00030442	0.006	
A*A	1	0.07237376	0.07237376	1.316	
S*I	1	0.02668832	0.02668832	0.488	
S*M	1	0.06472422	0.06472422	1.177	
S*A	1	0.08972965	0.08972965	1.632	
I*M	1	0.26356977	0.26356977	4.793	0.05
I*A	1	0.01460361	0.01460361	0.266	
M*A	1	0.02028663	0.02028663	0.369	
Residual	(35)	1.65735783	0.04735308		
LOF	10	0.28272549	0.02827255	0.514	
ERROR	25	1.37463234	0.05498529		
	49	3.08409221			

Table 9. ANOVA Summary for Total Search Time

SOURCE	DF	SS	MS	F	p ≤
Regression	(14)	6.79000303	0.48500022	8.821	0.005
Speech (S)	1	0.11508634	0.11508634	2.093	
Input time out (I)	1	5.58724900	5.58724900	101.600	0.005
Music Level (M)	1	0.23888390	0.23888390	4.345	0.050
Age (A)	1	0.29076850	0.29076850	5.288	0.050
S*S	1	0.00446760	0.00446760	0.081	
I*I	1	0.00005745	0.00005745	0.001	
M*M	1	0.00035065	0.00035065	0.006	
A*A	1	0.07316264	0.07316264	1.331	
S*I	1	0.02668832	0.02668832	0.485	
S*M	1	0.06472034	0.06472034	1.177	
S*A	1	0.08982066	0.08982066	1.634	
I*M	1	0.26387275	0.26387275	4.799	0.050
I*A	1	0.01461204	0.01461204	0.266	
M*A	1	0.02026284	0.02026284	0.369	
Residual	(35)	1.69161652	0.04833190		
LOF	10	.31698918	0.03169892	0.577	
ERROR	25	1.37462734	0.05498509		
	49	8.48161955			

In Table 9 the ANOVA summary table for total search time is presented. The regression model accounted for a large part of the total variance (R^2=0.8006). Input time-out ($F_{1,25}$=101.650, $p<0.05$), music level ($F_{1,25}$=4.345, $p<0.05$), age ($F_{1,25}$=5.288, $p<0.05$), and the interaction of input time-out by music level ($F_{1,25}$=4.799, $p<0.05$) were significant predictors. Lack-of-Fit was not significant ($F_{10,25}$=0.577, $p>0.1$).

Music level, age, and the input time-out by music level interaction were significant predictors for both total search time and subject search time. System time was solely predicted by the changes of input time-out values. Speech rate affected only subject search time but did not significantly influence total search time.

Figure 3 is a plot of total search time/trial versus the variable age in coded values. About 51 seconds (0.857 min.) of the total search time was contributed by the system. Computer simulations of the search tasks were used to generate the system times; therefore, the system time was constant and could not have been affected by the age of the subjects. The subject search time though, increased monotonically with age. Subject search time increased from about 4.8 seconds at the -1.414 coded level (15 year-old) to 19.2 seconds at the +1.414 coded level (60 year-old).

Figure 4 depicts how search time relates to input time-out. Input time-out was a significant predictor for both the dependent variables total search time and system time. However, subject search time was not significantly predicted by input time out. Figure 4 shows the linear nature of total search time. Subject search time contributed on the average a 10 second constant across the entire range of input time-out values. System times varied from 22.1 seconds at the -1.414 coded level (2 seconds of input time-out) to 80.8 seconds at the +1.414 coded level (10 seconds of input time-out).

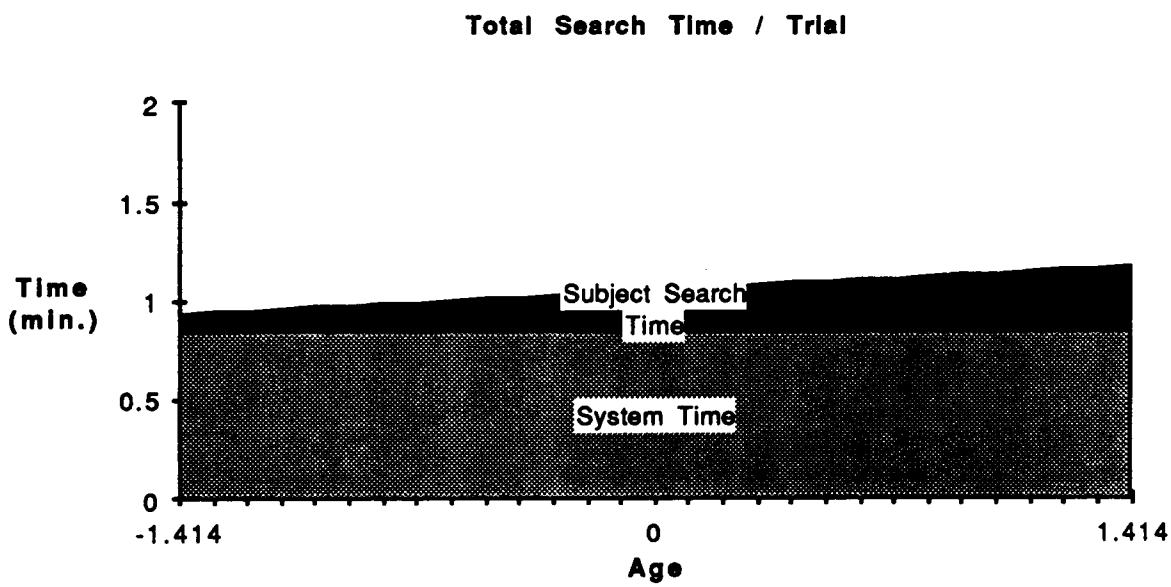


Figure 3. Total Search Time vs. Age

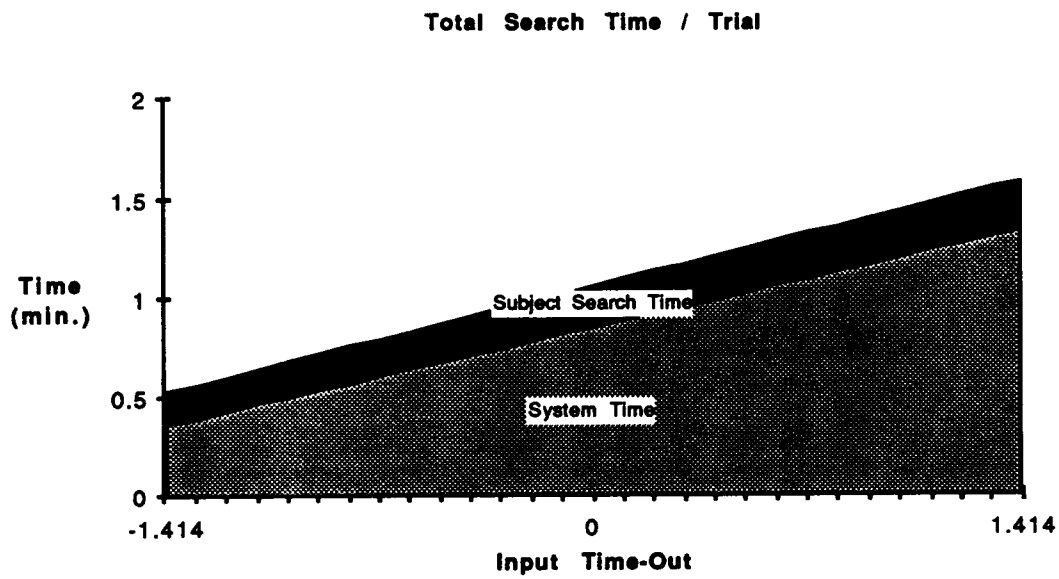


Figure 4. Total Search Time vs. Input Time-Out

Figure 5 illustrates the relationship between search time and music level. About 51 seconds of the total time was contributed by the system. The amount of time contributed by the system was exactly the same as it was in the case of variable age (0.857 min.). Again, one can not expect the music level would have any effects on the computer simulation and hence the exact same amount of system time was calculated. Search time increased with an increase of the level of music. Increase in background music level caused subjects to increase the amount of time it took to retrieve an item from 5.5 seconds at the -1.414 coded level (36.1 dB) to 18.5 seconds at the +1.414 coded level (65.4 dB).

Subject search time and the variable speech rate are presented on Figure 6. Speech rate was not a significant predictor of total search time, but it was a significant predictor for the subject contribution to search time. Hence Figure 6 plots only the subject search times against speech rate. Search times increased from approximately 4.5 seconds when speech rate was set at the -1.414 level (120 words per minute) to 19.5 seconds when speech rate was set at the +1.414 level (240 words per minute).

The interaction of input time-out by music level was a significant predictor for both total search-time and subject search-time. Figure 7 illustrates how system time varies with input time-out. System time increases from a minimum value of 0.37 minute at the -1.414 coded level (two seconds of time-out) to a maximum value of 1.35 minute at the + 1.414 coded level (10 seconds of time-out). The surface is flat because music level has no effect on the computer generated system time.

Figure 8 is the response surface plot of input time-out values and music levels against subject search time. Subject search time increased slightly with short input time-out (-1.414 at the coded level) and low music level (-1.414 at the coded level).

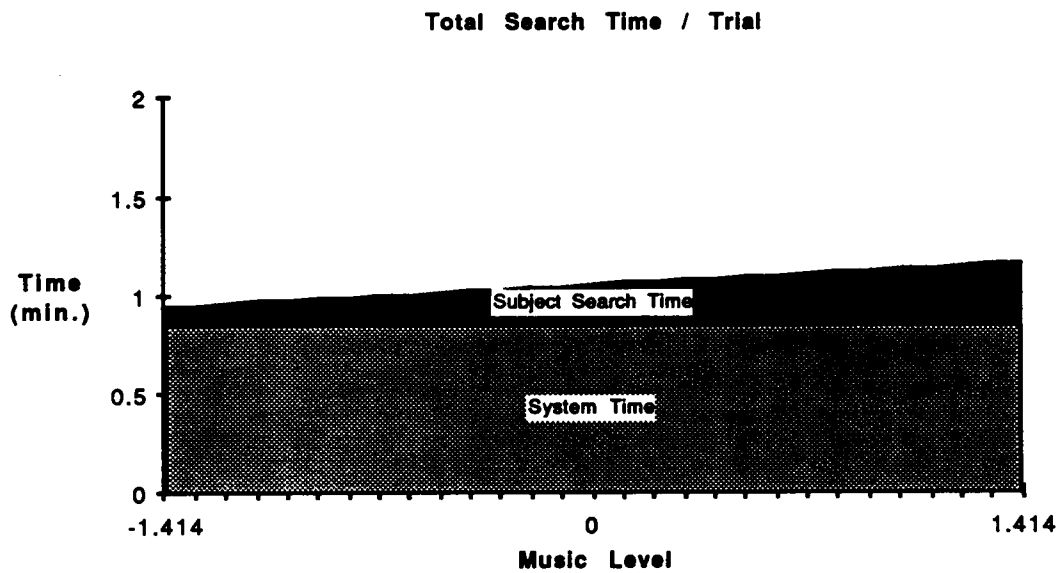


Figure 5. Search Time vs. Music Level

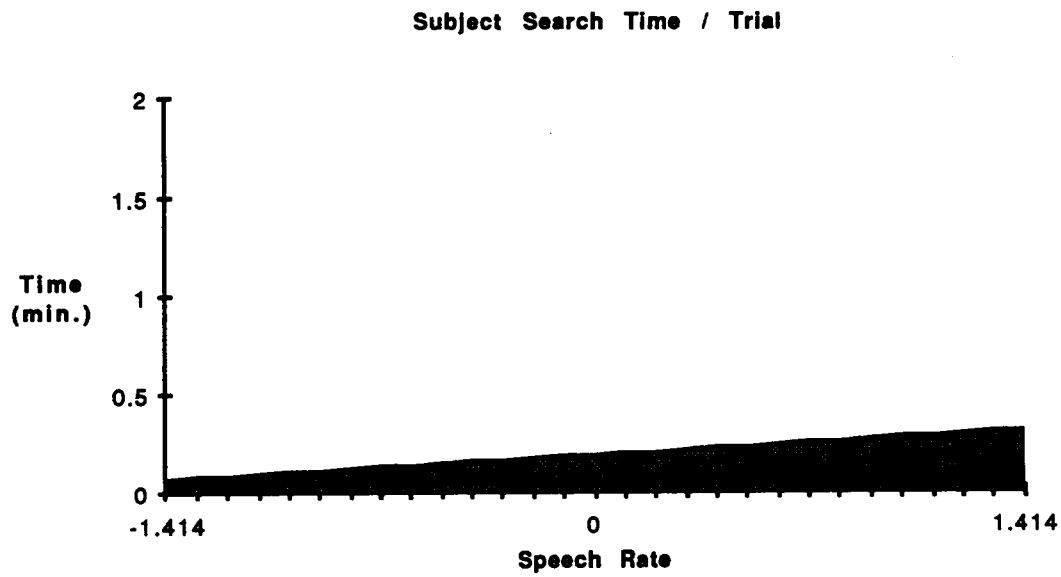


Figure 6. Subject Search time vs. Speech Rate

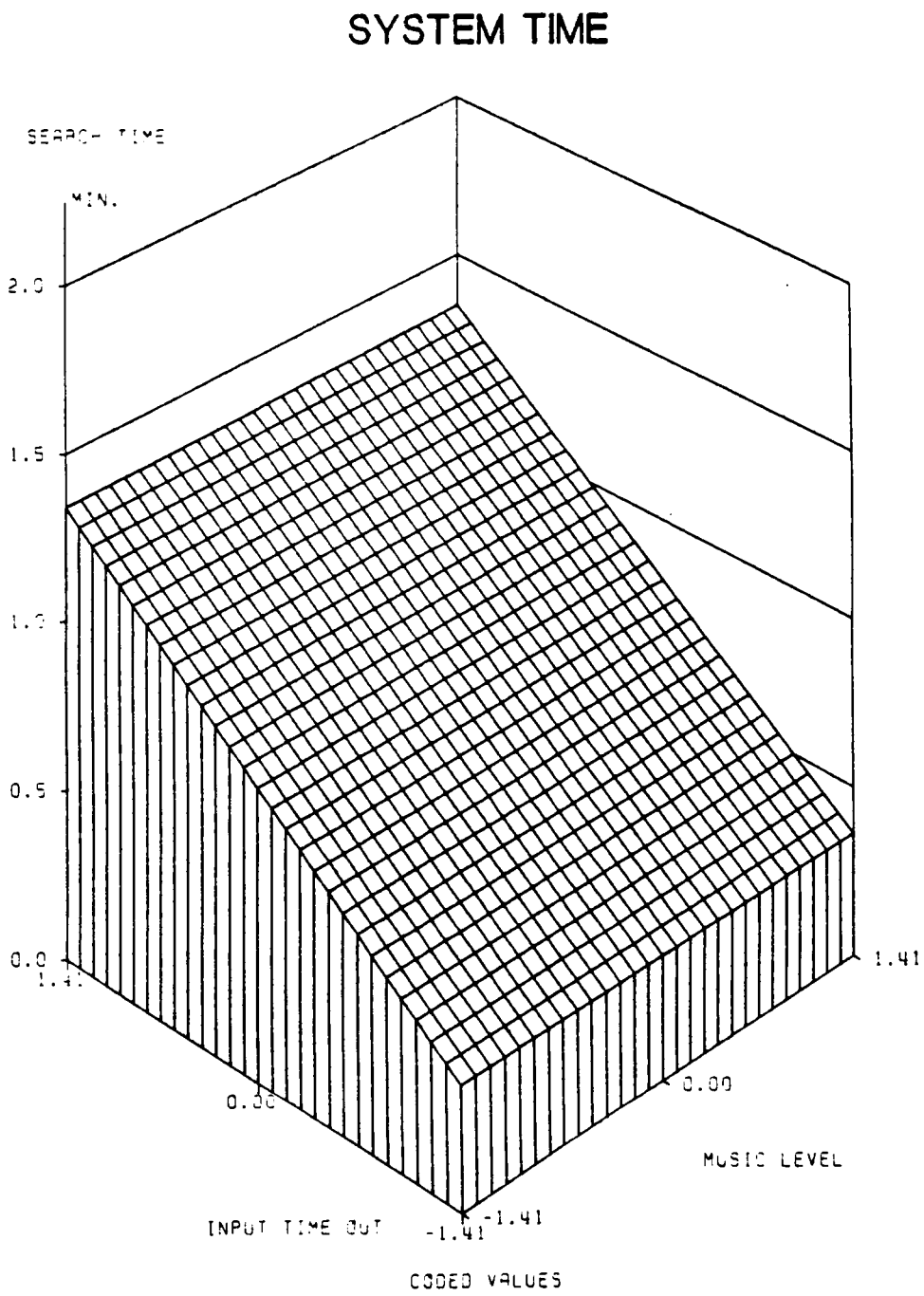


Figure 7. System Time for the Input Time-Out * Music Level Interaction

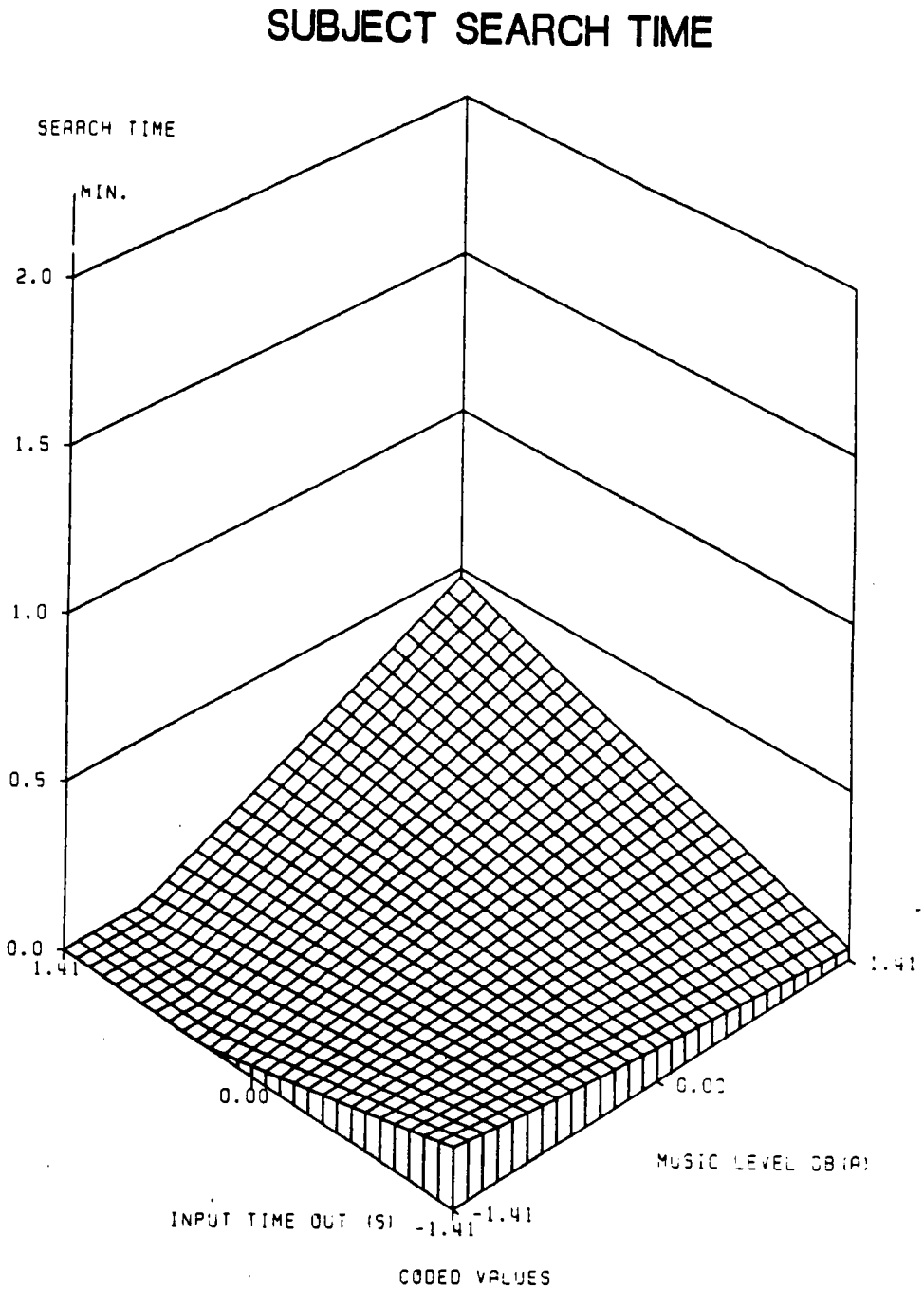


Figure 8. Subject Search Time for the Input Time-Out * Music Level Interaction

Search time increased as input time-out and music reached the 0 coded value with a very steep surface extending to near the 0.5 minute mark when both music level and input time-out reached the +1.414 coded level. Figure 9 represents the total search time which was created by adding subject search time to system time.

Keypresses

The number of extraneous keypresses made by a subject was selected as a measure of search error. A measure of extra keypresses was obtained by subtracting three (the minimum number of keypresses necessary to complete a trial) from the number of keypresses a subject made in each trial. The search portion of the experiment required at least two keypresses to reach a desired item, an additional keypress was required to enable the speech system to speak the information message. An average of 3.42 keypresses were made by the 50 subjects (σ^2=0.59). Fifteen subjects made no errors in the search portion of the experiment. One subject averaged as much as 5.94 keypresses per search.

Table 10 summarized the regression equation for the variable extra keypresses. The R^2 value for extra keypresses was 0.4331. Table 11 is the ANOVA summary table for extra keypresses. Music level was a significant predictor ($F_{1,25}$=4.839, $p<0.05$), and age was also significant ($F_{1,25}$=5.179, $p<0.05$). The Lack-of-Fit term was not significant ($F_{10,25}$=0.354, $p>0.25$).

Figure 10 illustrates the relationship between extra keypresses and age. Extra keypresses increases from 0.12/trial at the -1.414 coded level (15 year-old) to 0.71/trial at the +1.414 coded level (60 year old).

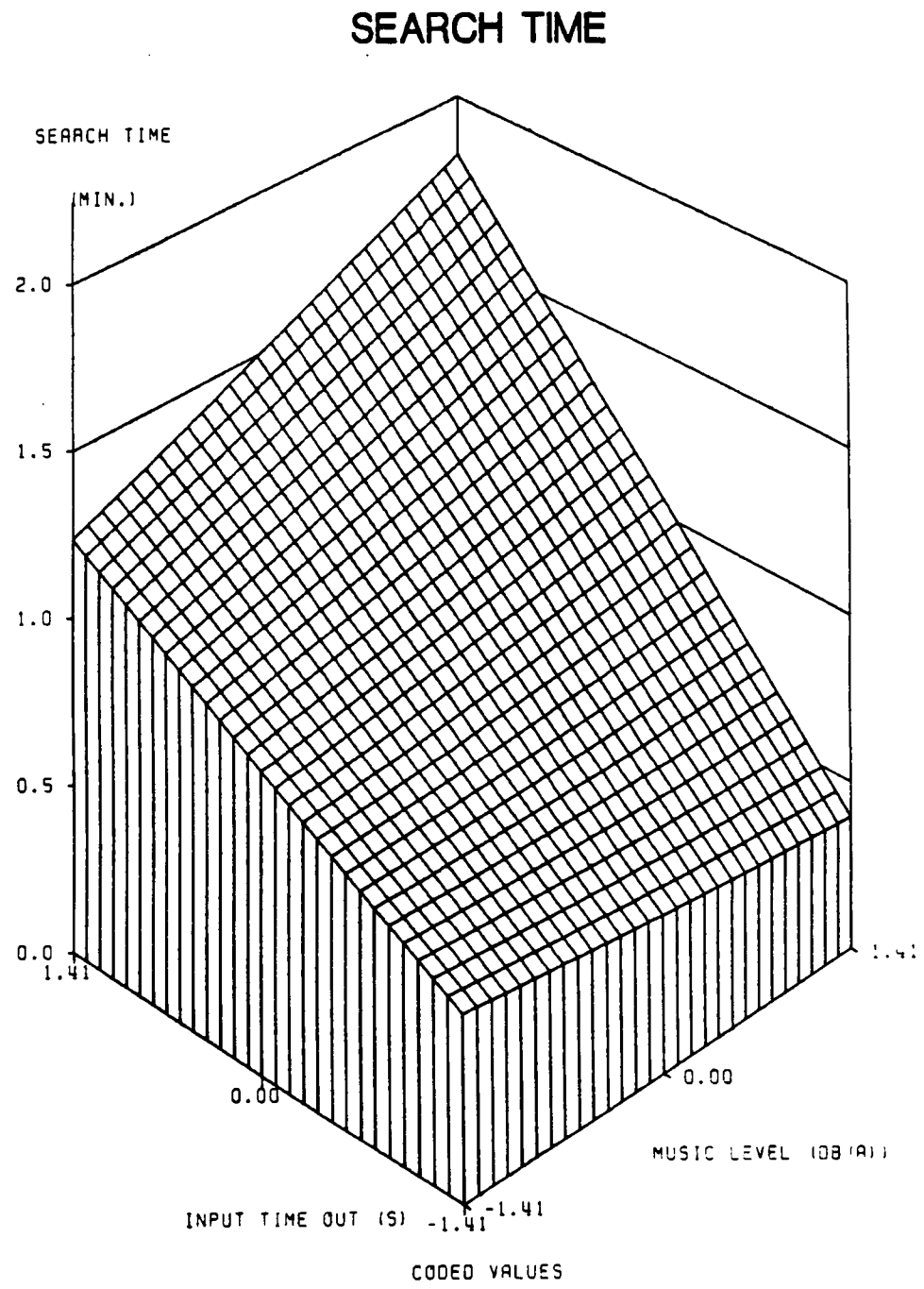


Figure 9. Total Search Time for the Input Time-Out * Music Level Interaction

Table 10. Regression Equations for Extra Keypresses

Variable	df	Estimate (β)	Std. Error	T ratio	Prob
Intercept	1	0.096	0.223	0.431	0.6689
Speech Rate (S)	**1**	**0.169**	**0.083**	**2.036**	**0.0494**
Input Time-Out (I)	1	-0.021	0.083	-0.248	0.8054
Music Level (M)	**1**	**0.203**	**0.083**	**2.436**	**0.0201**
Age (A)	**1**	**0.210**	**0.083**	**2.520**	**0.0164**
S^2	1	0.084	0.132	0.642	0.5253
I^2	1	0.100	0.132	0.760	0.4521
M^2	1	0.108	0.132	0.820	0.4179
A^2	1	0.108	0.132	0.820	0.4179
SI	1	-0.008	0.093	-0.084	0.9335
SM	1	0.039	0.093	0.420	0.6770
SA	1	0.125	0.093	1.344	0.1875
I M	**1**	**0.191**	**0.093**	**2.058**	**0.0471**
IA	1	-0.059	0.093	-0.630	0.5327
MA	1	0.105	0.093	1.134	0.2644

Table 11. ANOVA Summary for Extra Keypresses

SOURCE	DF	SS	MS	F	p≤
Regression	(14)	7.39799422	0.52842816	1.557	
Speech (S)	1	1.14719333	1.14719333	3.381	
Input time out (I)	1	0.01704968	0.01704968	0.05	
Music Level (M)	1	1.64168944	1.64168944	4.839	0.05
Age (A)	1	1.75723755	1.75723755	5.179	0.05
S*S	1	0.11390625	0.11390625	0.336	
I*I	1	0.16000000	0.16000000	0.472	
M*M	1	0.18597656	0.18597656	0.548	
A*A	1	0.18597656	0.18597656	0.548	
S*I	1	0.00195313	0.00195313	0.006	
S*M	1	0.04882813	0.04882813	0.144	
S*A	1	0.50000000	0.50000000	1.474	
I*M	1	1.17236328	1.17236328	3.455	
I*A	1	0.10986328	0.10986328	0.324	
M*A	1	0.35595703	0.35595703	1.049	
Residual	(35)	9.68489642	0.27671133		
LOF	10	1.20247454	0.12024745	0.354	
ERROR	25	8.48242188	0.33929688		
	49	17.08289063			

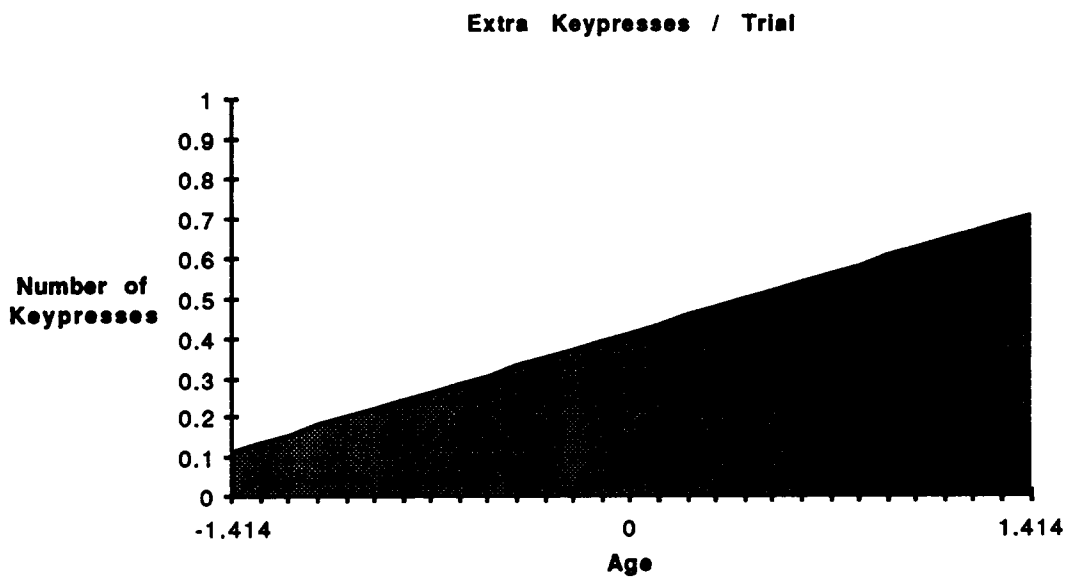


Figure 10. Extra Keypresses vs. Age

As music level increases, the number of erroneous keypresses also increases. At the -1.414 coded level (36.1 dB), the number of extra keypresses is 0.13. The number of erroneous keypresses increased to 0.70 with the music level set at the +1.414 coded level (65.4 dB). Figure 11 illustrates the relationship between erroneous keypresses and music level.

Transcription Errors

In the second part of the experiment, the subjects were required to transcribe information messages spoken by the speech synthesizer. Transcription error was a measure of how poorly subjects understood the information messages. Since only the first two and the last two words of each sentence were scored, the highest number of errors one could obtained was four words/trial. On average, error rate for the 50 subjects was 1.08 words/trial (s^2=0.615). The minimum error rate was 0.1875, which translates to three transcription errors for the entire search tasks (64 words). One subject's error rate was as high as 2.5 words/trial (40 words/experiment).

The regression equation summary for transcription error is presented in Table 12. The regression model has a R^2 value of 0.5724. Table 13 is the ANOVA summary table for the 14 predictors of transcription error. Speech rate ($F_{1,25}$=8.315, $p<0.01$), music level ($F_{1,25}$=27.303, $p<0.005$), age ($F_{1,25}$=5.615, $p<0.05$), and the pure quadratic term of music level - M^2 ($F_{1,25}$=6.513, $p<0.025$) were all significant predictors for the variable transcription error. The Lack-of-Fit term was not significant at the 0.05 level ($F_{10,25}$=1.809, $p>0.01$).

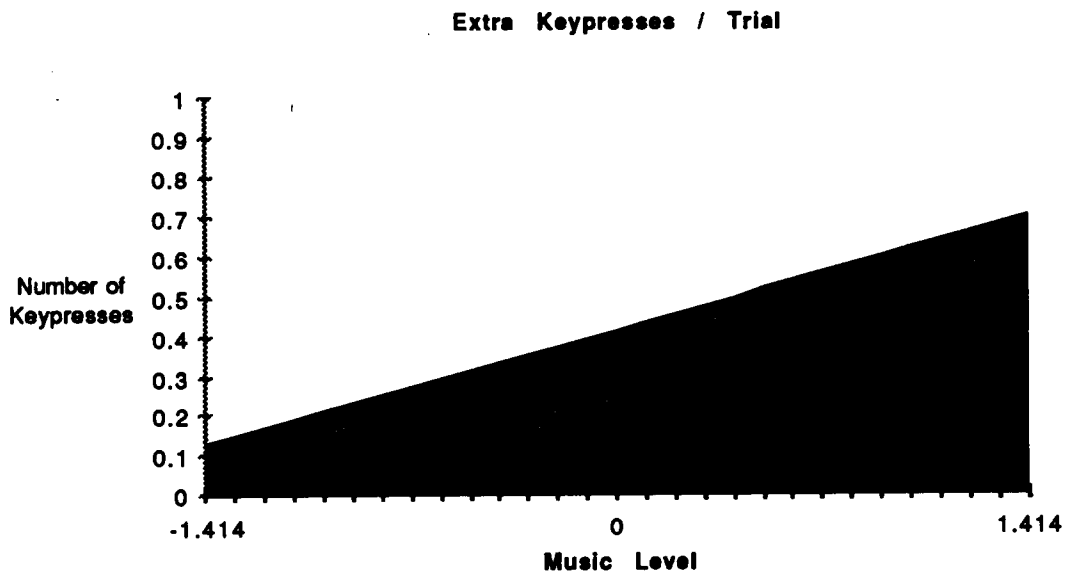


Figure 11. Extra Keypresses vs. Music Level

Table 12. Regression Equations for Transcription Errors.

Variable	df	Estimate (β)	Std. Error	T ratio	Prob
Intercept	1	0.831	0.202	4.120	0.0002
Speech Rate(S)	1	0.195	0.075	2.599	0.0136
Input Time-Out(I)	1	-0.014	0.075	-0.186	0.8539
Music Level (M)	1	0.354	0.075	4.709	0.0001
Age (A)	1	0.161	0.075	2.136	0.0398
S^2	1	0.047	0.119	0.394	0.6958
I^2	1	-0.133	0.119	-1.117	0.2715
M^2	1	0.273	0.119	2.300	0.0275
A^2	1	0.125	0.119	1.051	0.3003
SI	1	-0.055	0.084	-0.651	0.5196
SM	1	0.098	0.084	1.162	0.2532
SA	1	0.074	0.084	0.883	0.3833
IM	1	-0.004	0.084	-0.046	0.9632
IA	1	0.035	0.084	0.418	0.6784
MA	1	-0.141	0.084	-1.673	0.1033

Table 13. ANOVA Summary for Transcription Errors.

SOURCE	DF	SS	MS	F	p ≤
Regression	(14)	10.59491564	0.75677969	4.120	0.005
Speech Rate (S)	1	1.52719325	1.52719325	8.315	0.01
Input time out (I)	1	0.00778573	0.00778573	0.040	
Music Level (M)	1	5.01486040	5.01486040	27.303	0.005
Age (A)	1	1.03140439	1.03140439	5.615	0.05
S*S	1	0.03515625	0.03515625	0.191	
I*I	1	0.28222656	0.28222656	1.537	
M*M	1	1.19628906	1.19628906	6.513	0.025
A*A	1	0.25000000	0.25000000	1.361	
S*I	1	0.09570313	0.09570313	0.521	
S*M	1	0.30517578	0.30517578	1.662	
S*A	1	0.17626953	0.17626953	0.960	
I*M	1	0.00048828	0.00048828	0.003	
I*A	1	0.03955078	0.03955078	0.215	
M*A	1	0.63281250	0.63281250	3.445	
Residual	(35)	7.91484999	0.22613857		
LOF	10	3.32305311	0.33230531	1.809	
ERROR	25	4.59179688	0.18367188		
	49	18.50976563			

Figure 12 illustrates the relationship between transcription errors and age. Error rate increased from 0.85 words/trial at the -1.414 coded level (15 year old) to 1.31 words/ trial at the +1.414 coded level (60 year-old).

Transcription error varied in a curvilinear fashion with music level as shown in Figure 13. At the -1.414 coded level (36.1 dB), error rate was 0.91 words/trial. Performance was affected most severely in the range of 0 to +1.414 coded level. An error rate of 1.91 words/trial is obtained at the +1.414 coded level (65.4 dB). Error rate rose sharply from about the 0 coded level (50.8 dB) with a value of 0.86 words/trial to the +1.414 coded level (65.4 dB).

Speech rate was also a significant predictor for transcription errors. At the -1.414 coded level (120 words per minute), the error rate was 0.81 words/trial. Error rates increased to 1.36 words/trial at the +1.414 coded level (240 words per minute) as shown in Figure 14.

Location of Transcription Errors

Transcription errors were calculated by determining how many words were transcribed incorrectly by the subjects. As described in the method section, redundant prepositions and verbs were placed in the middle part of the sentences and were not scored. A paired t-test comparing each subject's transcription errors of the first two and last two words revealed significant differences between the locations of the words (t_{49} = -4.744, $p<0.0001$). Subjects made more errors in transcribing the first two words ($\bar{x}$ = 0.6661) than the last two words ($\bar{x}$ = 0.415).

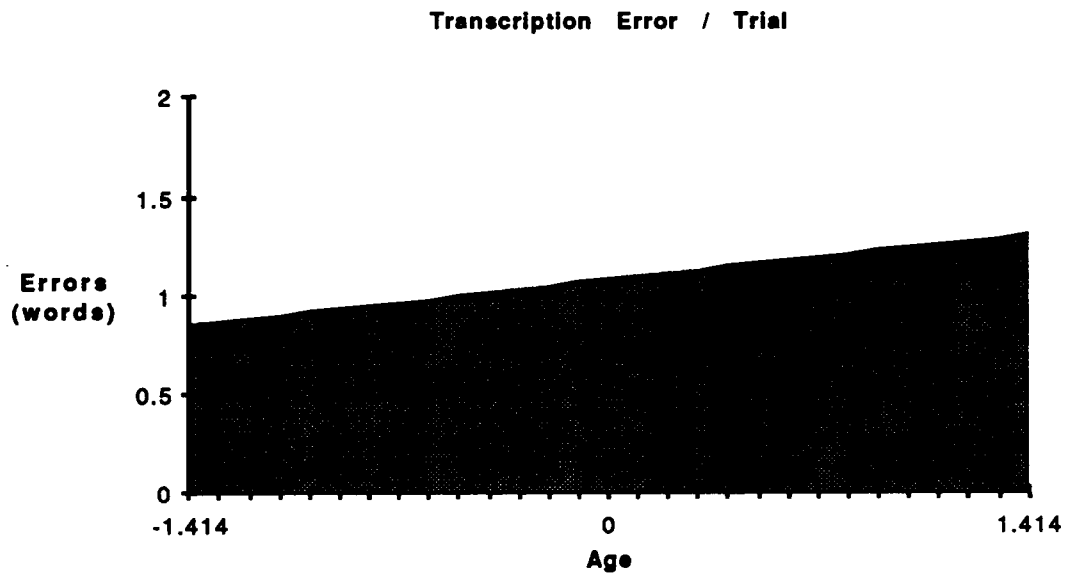


Figure 12. Transcription Errors vs. Age

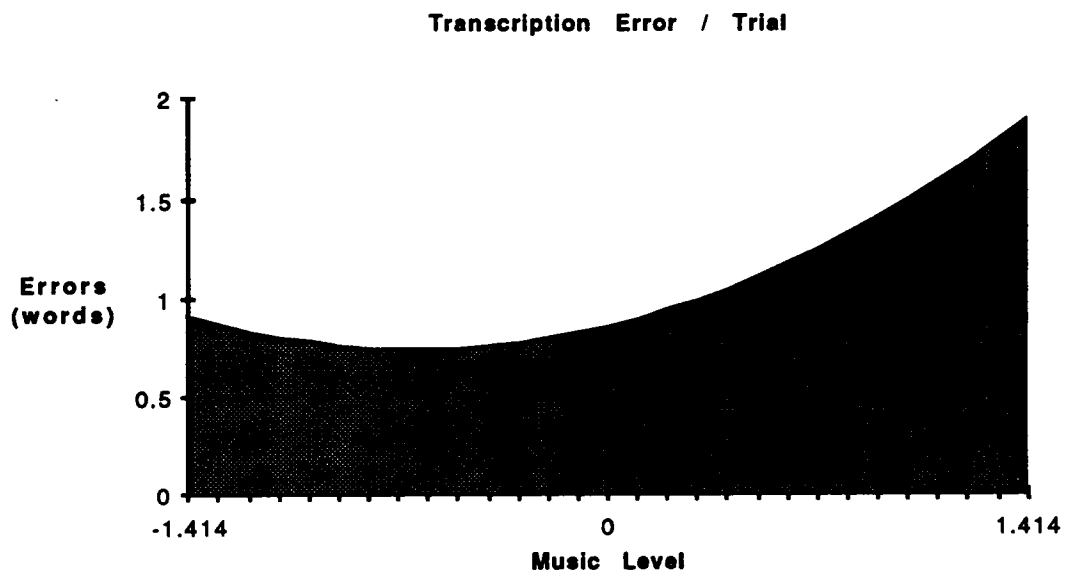


Figure 13. Transcription Errors vs. Music Level

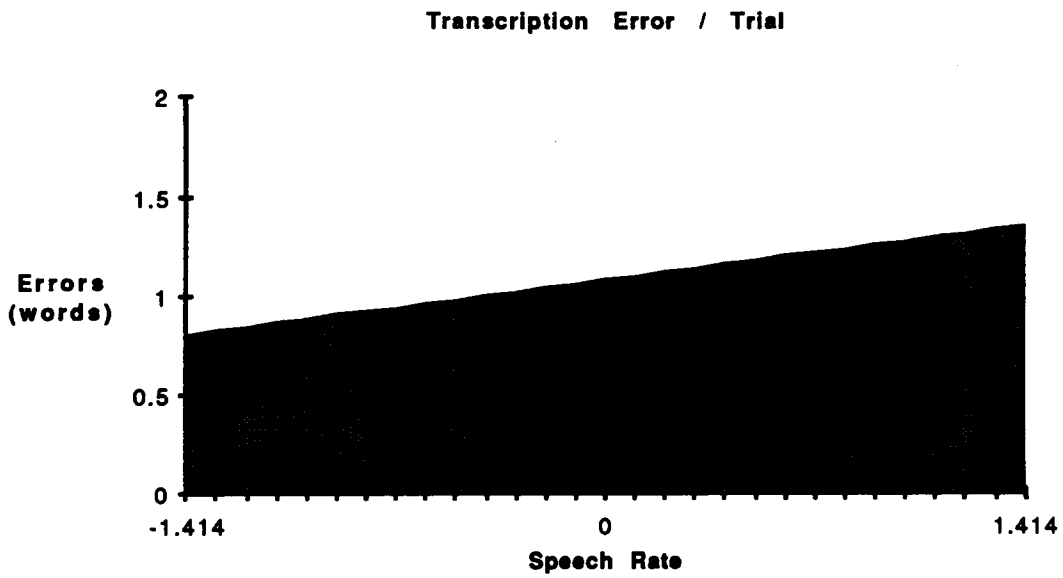


Figure 14. Transcription Errors vs. Speech Rate

Optimal Settings

The optimal conditions, presented in Table 14 were obtained by conducting the response surface regression procedure (Proc RSREG) in the SAS statistical analysis package. Each combination of four variables was calculated by obtaining four partial derivatives and setting the four derivatives to zero resulting in a combination of four unique factors. Some calculated settings might not be in the range of the tested values (i.e., value greater than +1.414 or smaller than -1.414); those values were obtained because in solving the four simultaneous equations, there was no range limitations on the values of the variables. The problem is quite apparent if one studies the settings of total search time; none of the four values were within the prescribed range and the translation to real-world values did not result in particularly meaningful settings. In calculating the optimal values, all eigen values of the optimal conditions had mixed signs, signifying that the calculated values were saddle points. Additional analysis such as canonical analysis should be conducted before conclusions can be drawn on whether the calculated values are representing true optimal values.

Subjective Ratings

Experimental Ratings

Each subject in the experiment performed 16 target searches; after each target search, each subject rated three quality measures of the particular search: certainty in transcribing the information message, difficulty of understanding the information message, and the difficulty of locating the target item. Therefore a total of 16 ratings were solicited from each subject in one experiment session. A median value of the 16 ratings was then calculated for each subject.

The rating data did not conform to the assumptions of parametric data analysis; for example, the data were not normally distributed and they were not measured on an

Table 14. Optimal Settings of Independent Variables.

	Subject Search Time		Total Search Time	
	coded value	real-world value	coded value	real-world value
Speech Rate	2.032	266.21 wpm	-37.61	-1415.83 wpm
Input Time-Out	-1.499	1.76 sec	14.45	46.88 sec
Music Level	-0.646	44.06 dB	4.67	99.10 dB
Age	-1.075	20.52 yrs.	10.98	216.64 yrs.
Optimal Values:	0.1313 min.		3.29 min.	

	Transcription Error		Extra Keypresses	
	coded value	real-world value	coded value	real-world value
Speech Rate	-0.593	154.84 wpm	-0.100	175.76 wpm
Input Time-Out	-0.040	5.89 sec.	-0.931	3.37 sec
Music Level	-0.772	42.75 dB	0.626	57.23 dB
Age	-0.895	23.44 yrs.	1.473	14.04 yrs.
Optimal Values:	0.5650 words		0.0064 keypresses	

interval scale. Therefore, the data were analyzed by converting them into ranks using a Kruskal-Wallis one way analysis of variance non parametric procedure as described in Siegel and Castellan (1988).

All ten categories of rating were analyzed according to the three independent measures of speech rate, input time-out, and music level. The hypothesis was to test if each of the five levels (-1.414, -1, 0 +1, +1.414) of the independent variables have affected the subject ratings significantly. However, the Kruskal-Wallis test specifically stated that there must be more than five data points for each level of independent variable; since the +1.414 and the -1.414 level of each variable only has two data points, these two levels were excluded in the analysis.

Figure 15 is a plot of how difficult it was to understand the information messages against the three different music levels. Subjective ratings were significantly different ($KW=8.28$, $p<0.02$) between the +1 coded level (61.1 dB(A)) and the -1 coded level (40.5 dB(A)). Subjects rated the +1 coded level (61.1 dB(A)) significantly more difficult to understand than the -1 coded level (40.5 dB(A)).

Figure 16 plots the ratings of how annoying the music level was against the different music levels used in the study. Subjects rated the music level was significantly more annoying ($KW=22.54$, $p<0.001$) at the +1 coded level (61.1 dB(A)) than at the -1 coded level (40.5 (dB(A)).

Figure 17 plots how subjects perceived the speed of the different speech rates. Subjects rated the +1 coded level (222 wpm) speech rate as significantly ($KW=8.325$, $p<0.02$) faster than the -1 coded level (138 wpm).

No other ratings were found to be significantly affected by speech rate, input time-out, or music level. Therefore, these data are represented descriptively by histograms plotting the frequency of the seven point rating scale. No statistical analyses

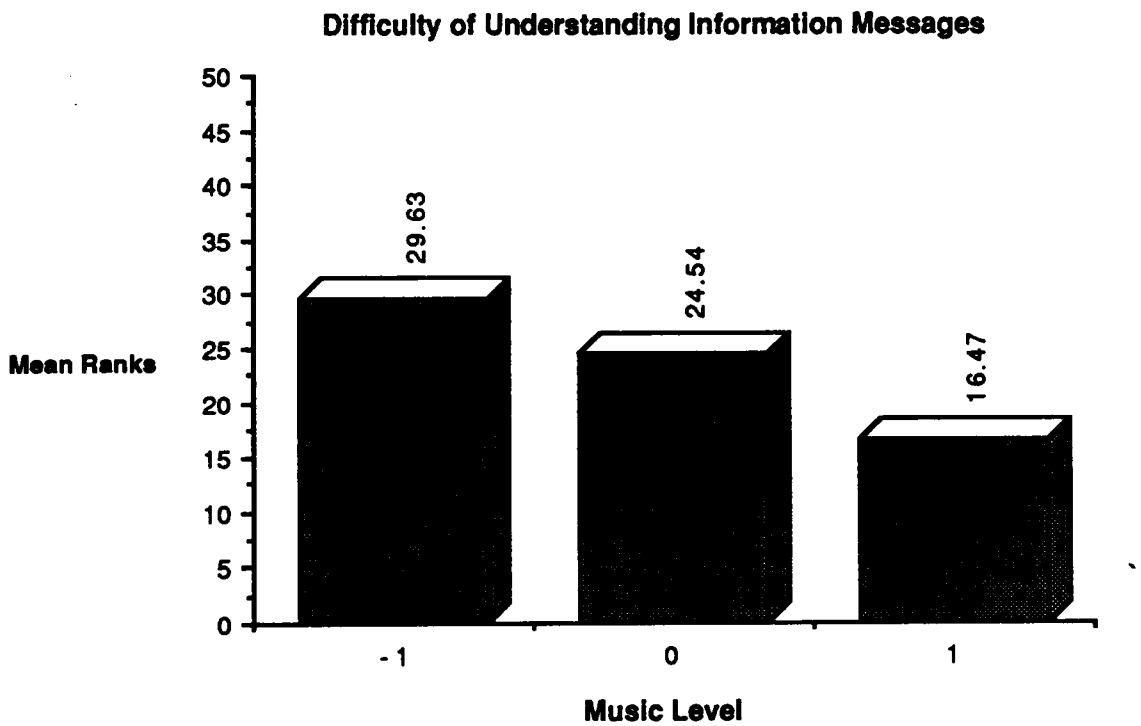


Figure 15. Difficulty of Understanding the Information Messages Ratings

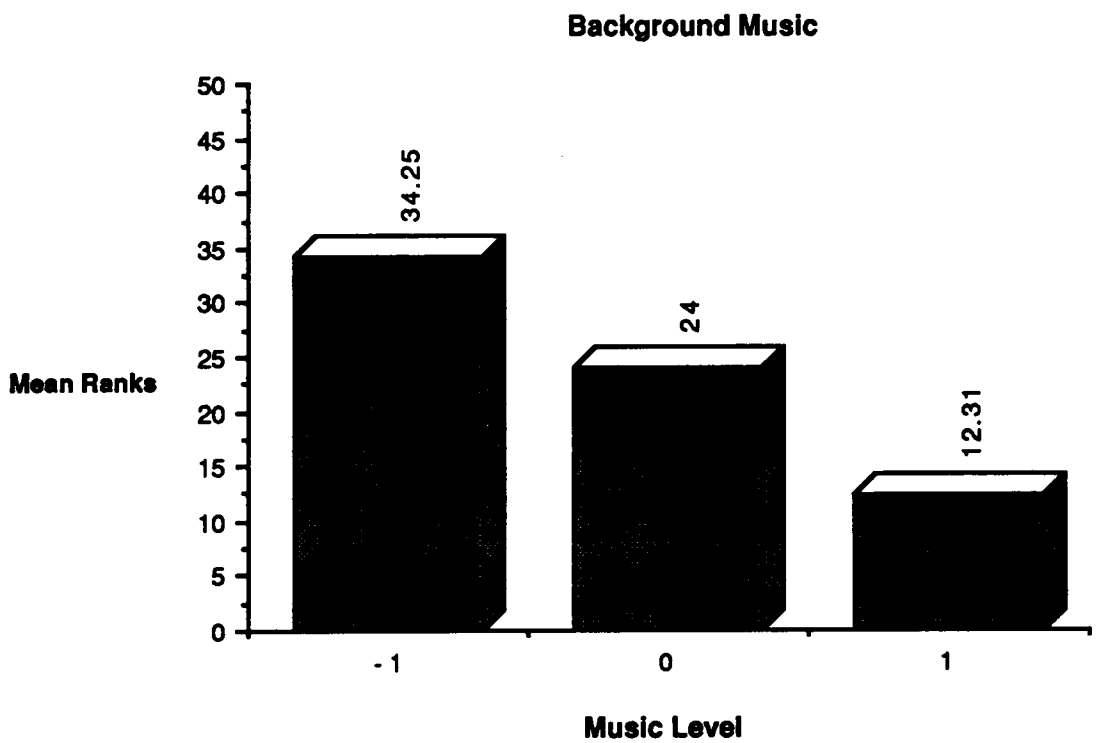


Figure 16. Background Music Level Ratings

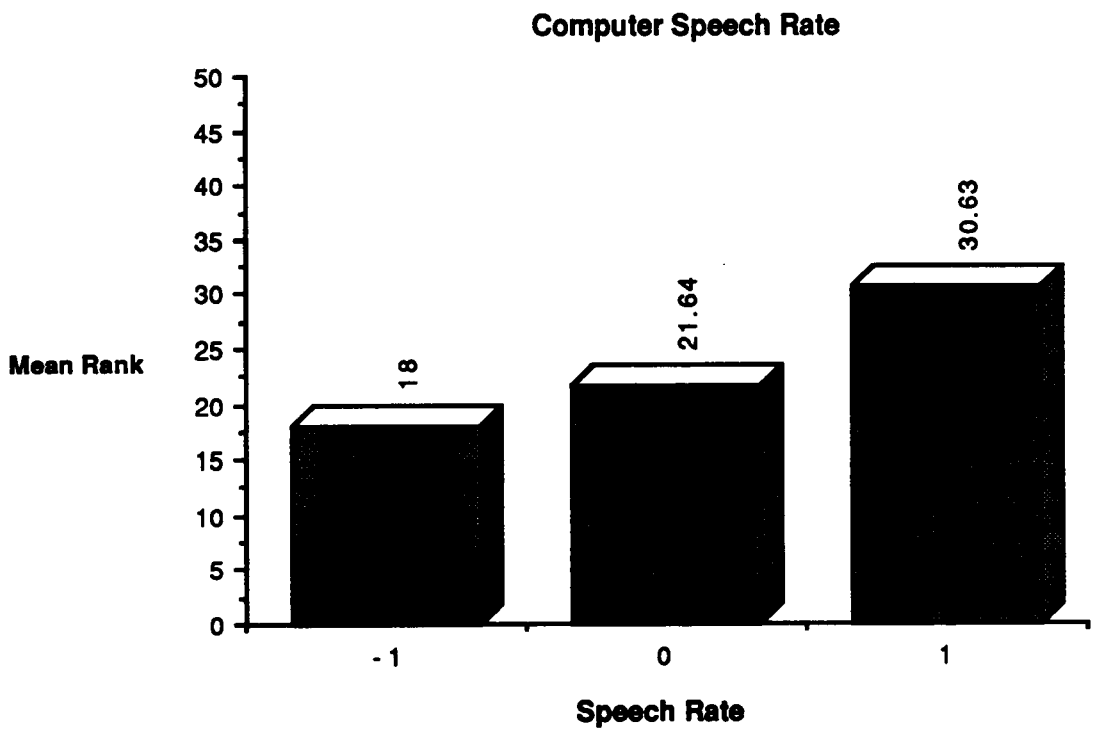


Figure 17. Computer Speech Rate Ratings

could be meaningfully conducted on these data because the frequency ratings of each point of the seven point scale were totally confounded with the treatment conditions.

In Figure 18, most subjects thought there was too much time for user input., the input time-out trend increased monotonically. No subject thought there was very little time for user input even though user input times were varied from two to 10 seconds.

Figure 19 plots the how certain the subjects were of their transcriptions. Thirty-six subjects rated the certainty at 4 or better.

Figure 20 represents how difficult it was to locate the target item in the search tasks. Forty-five subjects rated the difficulty as 6 or 7 indicating it was fairly easy to locate the target items.

The voice intelligibility rating frequencies were spread more evenly across the seven point scale. Eight people rated the intelligibility at 2 with an additional eight subjects rated it at 3. The ratings of 4 and 5 have frequency counts of 14 and 15, respectively. Only 4 people rated the intelligibility as high as 6 while no subject thought the intelligibility was good enough to deserve a 7 rating. The rating frequencies are presented in Figure 21.

Computer voice naturalness ratings are presented in Figure 22. As in the case of intelligibility, subjects did not give a rating of 7 for naturalness.

The frequency of the ease-of-use ratings were summarized in Figure 23. Fifteen subjects gave a rating of 5 while 43 of the 50 subjects gave ratings of 4 or better.

Most people described the menu as very organized as seen in Figure 24. Twenty-one subject rated the menu organization at 7 while no subject gave ratings below 3.

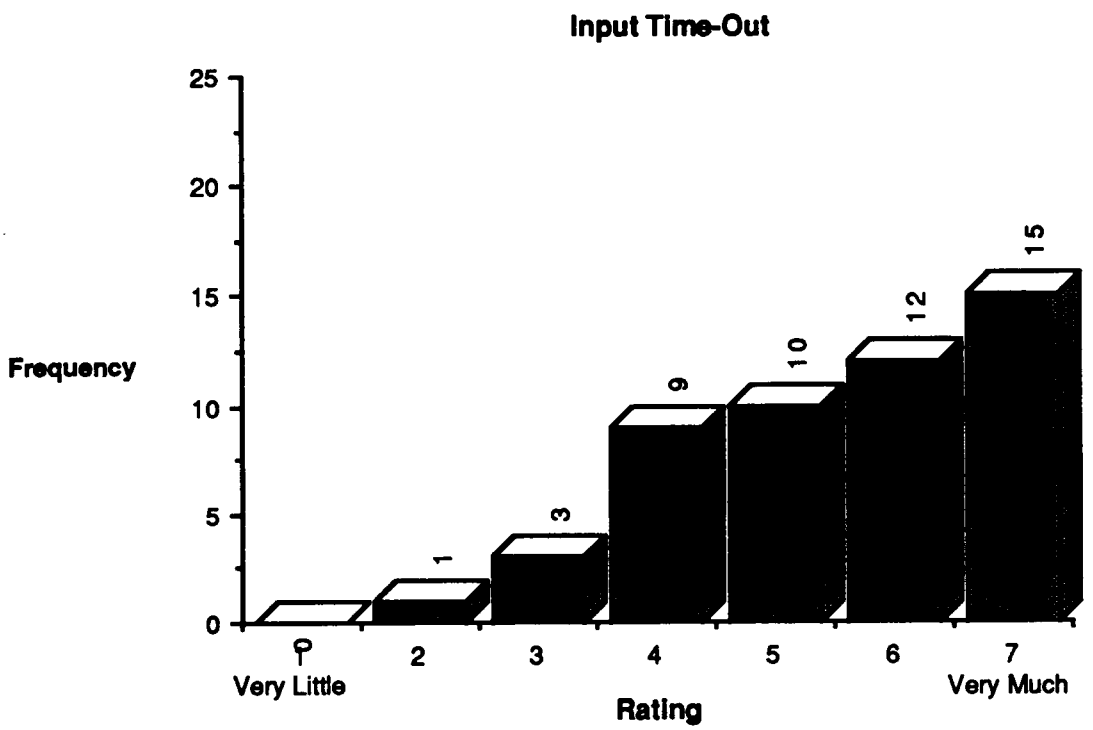


Figure 18. Input Time-Out Ratings

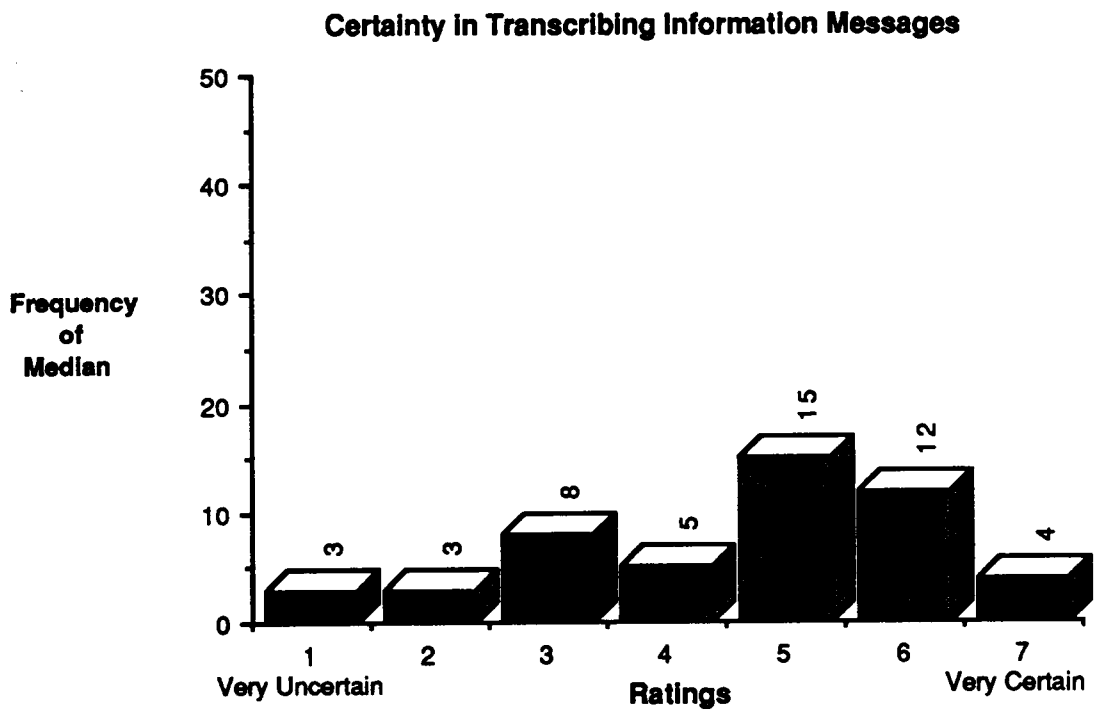


Figure 19. Certainty Ratings in Transcribing Messages

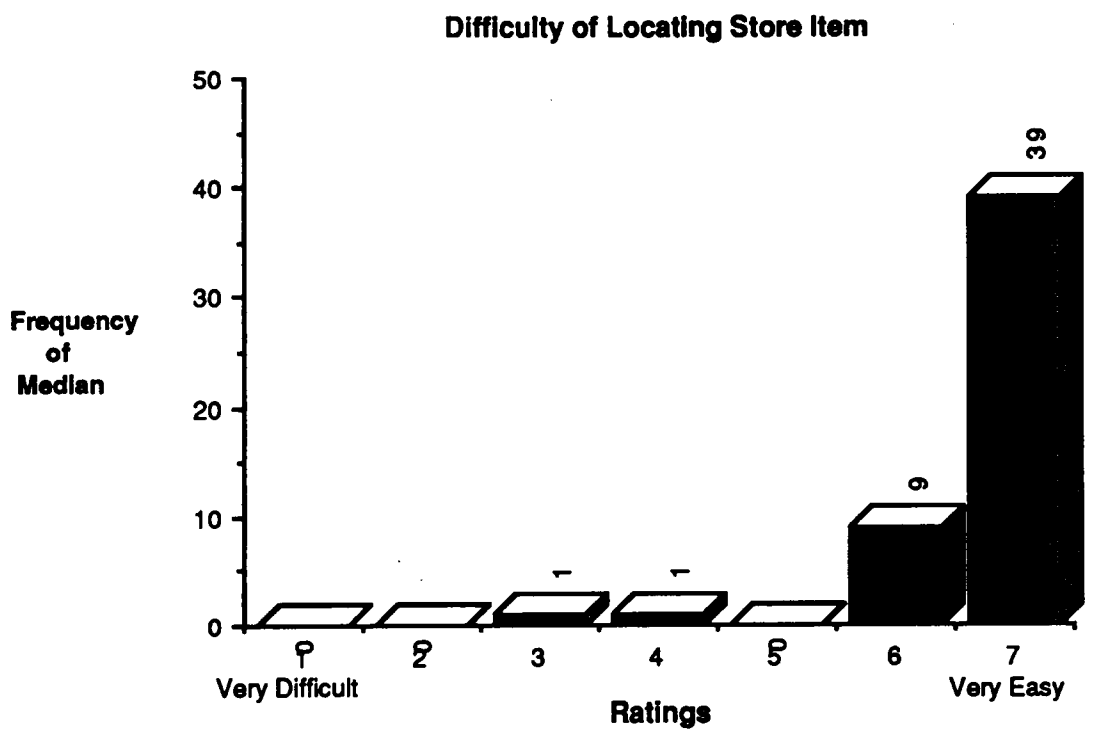


Figure 20. Difficulty Ratings in Locating Target Items

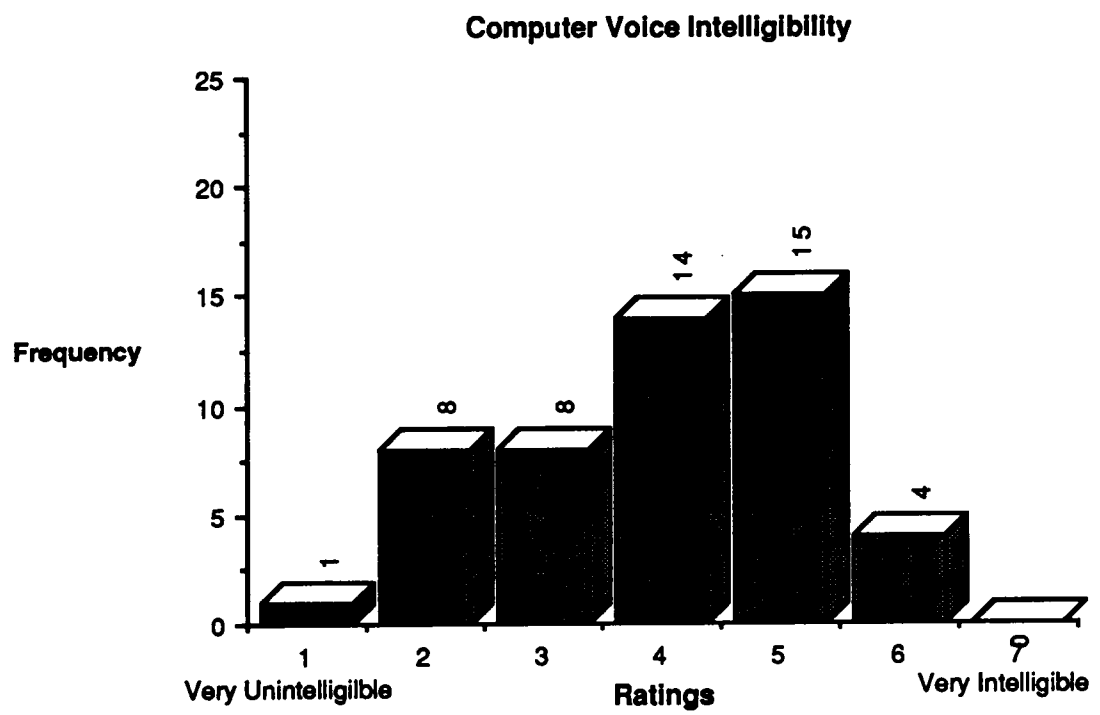


Figure 21. Intelligibility Ratings

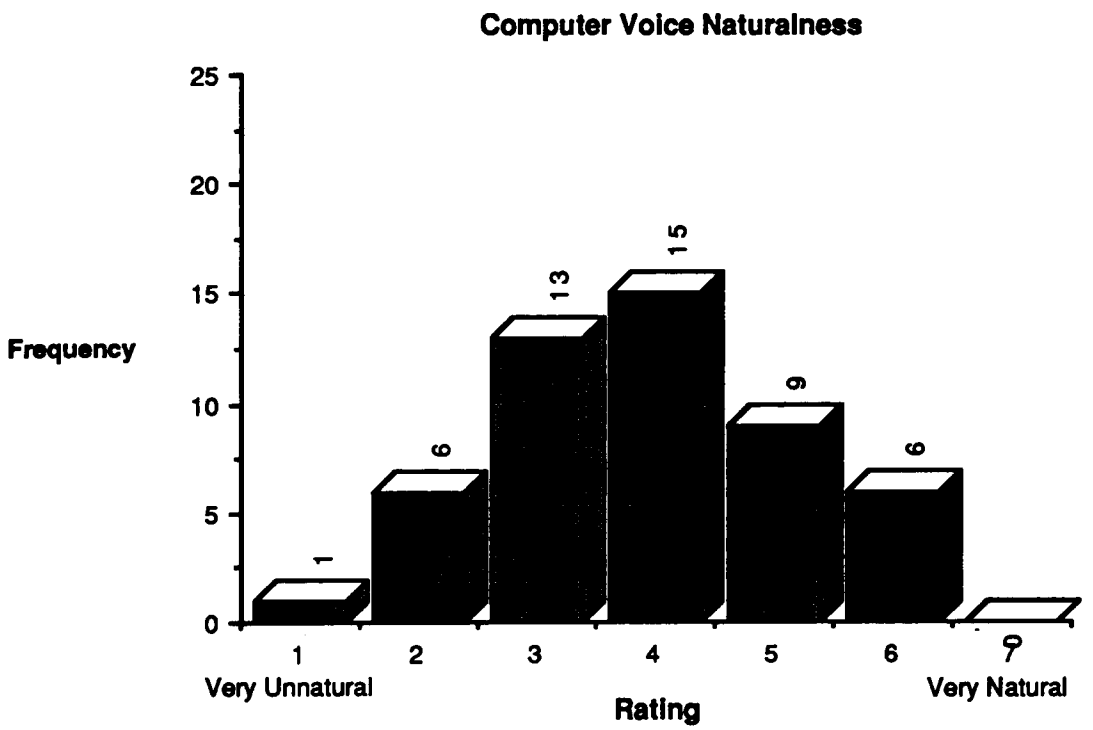


Figure 22. Naturalness of Computer Voice Ratings

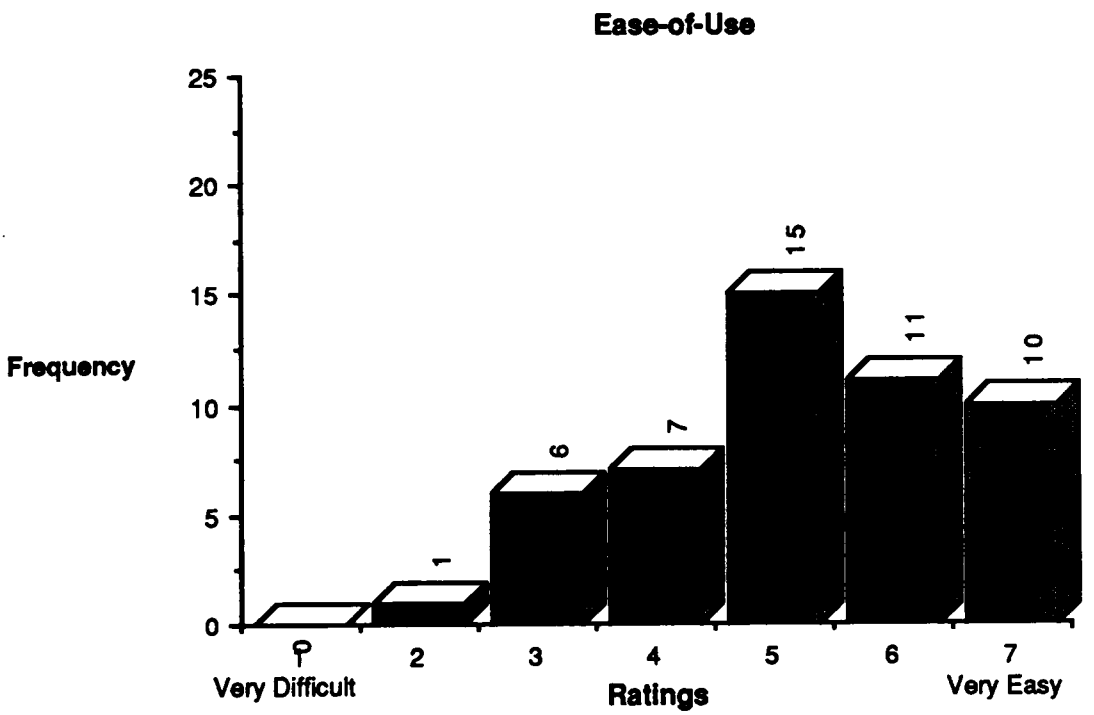


Figure 23. Ease of Use Ratings

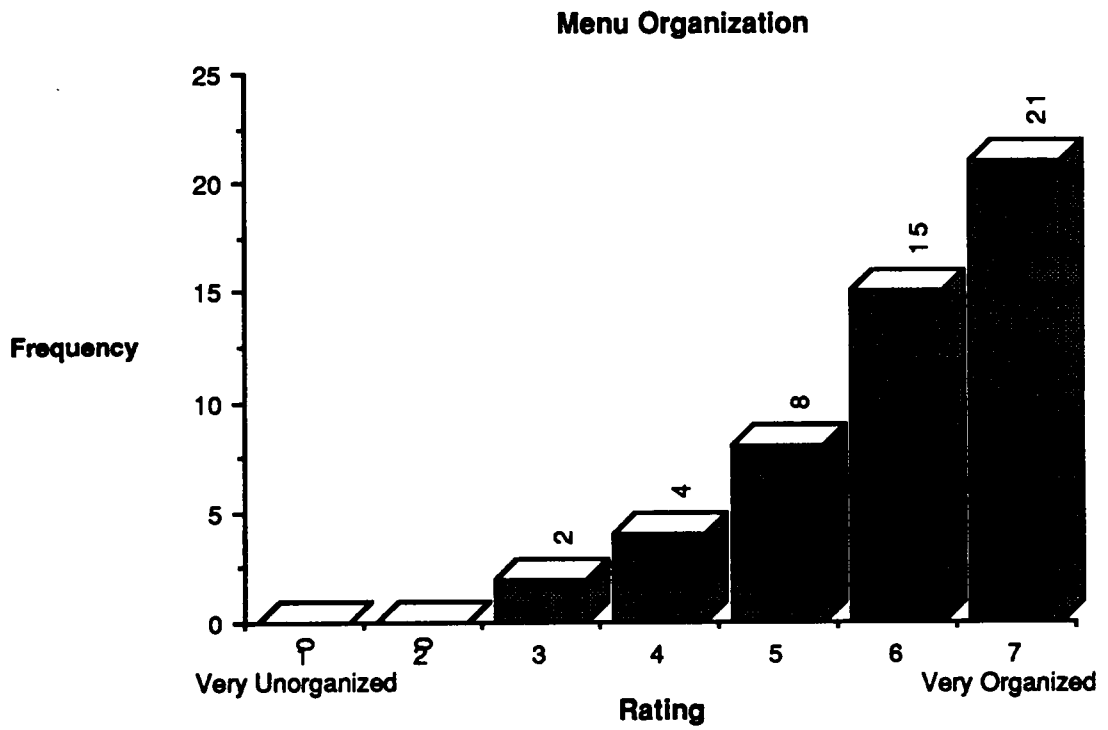


Figure 24. Menu Organization Ratings

DISCUSSION

The effects of the independent variables are discussed in this section, along with a number of important design and methodological issues.

Input Time-Out

Input time-out was a significant predictor for both system time and total search time. System time was almost entirely affected by input time-out because 99% of the variance of system time was accounted by input time-out. The system time was calculated by simulating the entire experiment to obtain estimates of how much time the system took to locate the item (mainly affected only by input time-out) and how much time the speech synthesizer took to speak the various keywords and information messages. Speech rate was not a significant factor in search time because only 1 % of the system time was accounted for by speech rate. From the regression equation, it was discovered that total search time varied linearly with input time-out. It is not surprising to see that increases in input time-out values would lead to increases in total search time, while system time was almost entirely affected by input time-out. From the results, search time was a constant 10 seconds/trial regardless of the input time-out level. However, the short input time-out value must reach a point where search time would increase with that setting of input time-out. For example, a short input time-out of 0.1 seconds would not allow a user enough time to press a key, let alone making the appropriate selection decision. Therefore, search time might be infinitely high at that level. The idea of studying input time-out was to select an optimal setting such that user would have enough time to make a decision and to press the correct key on

the keypad. From the regression equation for subject search time, search time was decreasing even at the two second mark. Therefore, one can conclude that two seconds was plenty of time for the user to accomplish the selection task. Subjective ratings also supported the hypothesis that there was plenty of time for user input as 46 of the 50 users responded by a rating of four or better. An input time-out setting less than two seconds might produce even better results.

Background Music Level

From the ratings data, 17 subjects rated the music level as very disruptive. In the objective measures, music level also affected subject and system search time and was a significant predictor in affecting the time required to retrieve the information. Music was loud enough to cause the subjects to incorporate a different strategy when they were listening to the synthesized voice. At conditions above +1 coded level (61.1 dB(A)) which were considered loud by most subjects, subjects adopted strategies such as putting their head very close to the speaker in order to listen to the speech system and covering one ear with one hand and put the other ear right on the speaker of the telephone. All participants when subjected to music level with 61.1 dB(A) or more, commented that the music were much too loud when asked by the experimenter in the debriefing period. However, from the debriefing notes, no subjects subjected to a coded value of -1 (40.5 dB) or lower complained about the music. At the 0 coded level (50.8 dB (A)), subjects' expressed mixed opinions about the disruptiveness of the music. Some subjects complained that the music was a bit too loud while some others echoed the sentiment that the music level was set at the correct level.

The increased level of music caused search time to increase. In most of the trials with music set at loud levels (> 61.1 dB(A)), subjects did not select the correct

keywords. The loud music did not enable the subjects to hear the keywords clearly, and consequently subjects employed a strategy such as listening to the complete menu (eight items) before making a selection; therefore lengthening the search for an item.

Transcription errors varied with music level in a non-linear fashion. Transcription errors rose sharply from the 0 coded level (50.8 dB(A)) to the +1.414 coded level (65.4 dB(A)). Music definitely had an effect on transcription abilities as some participants under loud music levels often guessed the information message when they were prompted to transcribe the information messages. It is also very interesting to note that transcription errors were higher when there was no music in the background. More transcription errors were calculated in the -1.414 coded level (36.1 dB(A)) than in the 0 coded level (50.8 dB(A)). The smaller number of transcription errors in a louder condition might be an artifact created by the subjects.

From the prediction equation, as the level of music increased, the number of keypresses subjects made also increased. The increase of music level often made some of the keywords difficult to understand. For example, when subjects were prompted to search for the item "feminine oriental fragrances". Some subjects selected "feminine floral fragrances" and were told that it was not the correct item, then the subjects had to back up or restart which lengthened the search time and increased the number of keypresses. Searches with lower background music often allowed subjects to distinguish between words such as "floral" and "oriental".

According to the rating data, subjects found that it was more difficult to understand the information messages at the +1 coded level of music (61.1 dB(A)) than at the -1 coded level (40.5 dB(A)). The ratings also revealed that the music was more annoying at the +1 coded level (61.1 dB(A)).

From the transcription error data, subjects certainly were more comfortable with music level set at or below 50.8 dB(A). Therefore, one should try to avoid background noise levels at or above 50 dB(A). Since the speech system used a speaker phone, background sound level might be higher before performance starts to drop off if the tasks were to be conducted using the telephone handset. Beranek et al. (1971) presented a guideline for maximum allowable background noise for certain activities. An appropriate background sound level of 47 to 56 dB constituted what they termed "fair listening conditions". Between background sound level of 47 to 56 dB, it was reported as: "suitable for lobbies, laboratory, work spaces, drafting and engineering rooms, and general secretary areas". Since background noise in this case was simulated by music, it might have varied a great deal more than steady background noise although the music was measured using the equivalent noise level (L_{eq}) method.

Speech Rate

Speech rate was a significant predictor for subject search time, but it was not a significant predictor of total search time. In the results section, it was shown that subject search time varied from 2.8 seconds/trial to 17.7 seconds/trial. In this study, only 1 percent of the system time variation was accounted for by speech rate. Therefore, even if speech rate is slow, it would not lengthened the search time by any noticeable magnitude. The variance in system time was explained mostly by the changes in input time-out. However, speech rate did increase the user's contribution to search time. Subject search time increased linearly with speech rate; therefore, one can conclude that by decreasing the speech rate, subject search time would decrease. Since speech rate was not a significant predictor of total search time and the speech rate by input time-out interaction was not significant, it is logical to conclude that manipulation of the speech rate would not have any dramatic effects on the total search time. However,

since system time was a major contributor to the prediction of total search time, it is possible that the changes in speech rate did not have a large enough effect to influence total search time. If input time-out values were smaller (e.g., 1 to 3 seconds), then speech rate might have played a more important role in determining the total search time.

Transcription errors increased as speech rates increased. From the plot of transcription vs. speech rate, error rates increased linearly with speech rate. Therefore, the slower the speech rate, the better one would be able to understand synthetic speech. From the regression equation, it was predicted that error rate would increased from 20% (0.81 words/ 4 words per trial) at the -1.414 coded level (120 wpm) to about 40% at the +1.414 (240 wpm) coded level.

Some researchers have concluded that speech rate of about 150 wpm is optimal in most transcription tasks (Slowiaczek and Nusbaum, 1985; Waterworth and Lo, 1984), while Merva (1987) observed that the optimal speech rate might lie above 150 wpm but below 180 wpm. The optimal settings calculated for the transcription task in this experiment did in fact suggested a speech rate of -0.593 coded value (155 wpm) as the optimal speech rate.

However, the transcription tasks were conducted with conditions such as loud background music or with older subjects. Therefore, the poor performance might have been the result of a combination of the other variables. Subjective ratings show that the users perceived the +1 coded level (222 wpm) as a faster speech rate than the -1 coded level (138 wpm).

In this study, most of the search time was contributed by input time-out. Using slow speech rates did not increase total system time greatly. However, slower speech rates did decrease transcription errors and decrease the subject's contribution to search

time. Therefore, speech rates at or below 150 wpm might be appropriate in the design of similar speech systems. The lowest rate tested was 120 wpm in this study. However, rates less than 120 wpm might be so slow that it would drastically lengthened the time to listen to the messages.

Age

The variable age was a significant predictor for all dependent measures. Performance, in most cases, worsened with the increase of age. In this study, the two youngest subjects both performed the task extremely well. Both subjects' search times were below the average at 0.15 and 0.06 minute as compared to 0.20 minute for the average subject. Their average extra keypresses of 0.1875 and 0.1250 were also less than that of the average (0.416). Their transcription errors were 0.8125 and 0.6875 words/trial as compared with the average of 1.0813 words/trial. Their results were an indication that subjects at that age were clearly capable of using telephone information system effectively. Conversely, one of the 60 year-old subject took over 0.805 minute to conduct one search while making an average of 2.5 words of error per trial. Contrary to the findings of the Waterworth and Lo (1984) study; which employed subjects with similar ages (14 to 64), this study show considerably difference in performance with the different age groups.

Input Time-Out by Music Level Interaction

The input time-out by music level interaction was a significant predictor of both subject and total search time. However, search time did not vary with input time-out. However, from the response surface, subject search time increased dramatically when music and input time-out was increased beyond the 0 coded level for both variables. There was a sharp increase in search time, especially with the combination of long

time-out periods and loud music. This increase can be explained by behavior during the experiment. With short input time-out periods, most subjects placed their fingers near the "#" button which was the button for selecting a menu item. Subjects were very eager and alert with short time-out values, fearing that they would missed the opportunity to select the appropriate item because once the subject passed on an item, eight keywords would be spoken before the same keyword appeared again. However, in longer time-out values, subjects were relaxed and did not display the same degree of alertness as their counterparts with short input time-out values. Because the subjects knew the time-out period was long, they were relaxed without putting their finger near the keypad. In the debriefing, some subjects complained about the music had distracted them from the search tasks; therefore, taking longer time to press the telephone keys.

Context Location of Information

Merva (1987) conducted an experiment involving information transcription and concluded subjects were more error-prone in the beginning part of the transcription task. This study obtained similar results with subjects transcribing the last two words more accurately than the beginning two words ($t_{1,49}$=-4.744; $p<0.0001$). The information messages of this study contained approximately six to eight items of information. It is well documented that human short term memory was discovered to hold approximately five to nine items of information, with distractions such as background music, it is not surprising that most subjects remembered the latter part of the information message better. Therefore, it is quite essential to save important contextual information for the end of a message.

Design Issues

Selection of Dependent Variables

In determining search times, speech rate was a significant predictor for subject search time but was not significant for total search time. Therefore, if the data were analyzed using only total search time, the effects of speech rate would have been lost. The selection of an appropriate dependent variable is, therefore, quite an important issue. First, since input time-out accounted for 99% of the variance in system time, it is very doubtful that changing the speech rate would have affected system time. Speech rate yielded a sums of squares of 0.3115 in the ANOVA summary table for subject search time, and the value decreased to 0.1151 in the ANOVA summary for total search time. Error terms in both tables were almost identical at 0.055. The sums of squares value of speech rate decreased drastically causing speech rate to be calculated as non-significant in the case of total search time. However, if one is to study the effects of human performance in the system; measures such as subject search time, not total search time, may be more appropriate.

Orthogonality of β weights

This experiment was formulated as an orthogonal design. In an orthogonal central-composite design, β weights across predictors should have no correlation and their variance should be purely additive.

One way to verify whether the variance is orthogonal in any experiment is to compare the Type I and Type II sums of squares obtained by running the regression procedure (Proc Reg) in the statistical analysis package SAS (SAS, 1985). Type I sums of squares are variances (sums of squares) that are order dependent; that is, each effect is adjusted only for the preceding effects in the model. For example, if the analysis is run with the higher order effects entered first, the resultant sums of squares would be

different than if the model is initially entered with first order effects. Type II sums of squares are calculated by assuming each effect is the last one entered into the model and hence are not order dependent. In an orthogonal design, Type I and Type II sums of squares should be identical.

Williges (1981) described that in a second-order, central-composite design; the number of data points are calculated by the Equation:

$$Q=[(F+2K+C)^{1/2}-F^{1/2}]^2 .$$

For this study, Q was calculated as equal to 1. The α value of the design is calculated by $\alpha=(\frac{QF}{4})^{1/4} \Rightarrow \alpha=(\frac{1(16)}{4})^{1/4} = \sqrt{2}$. The number of decimal places that was used in the data analysis for the α value caused certain β estimates to have different Type I and Type II sums of squares. Table 15 is a condensed summary table from the SAS output using the regression procedure (Proc Reg) with the variable subject search time using an α value of 1.414 with 3 decimal digits. The Type I and Type II sums of squares of the regressors S^2, I^2, and M^2, were different. It is an indication of non-orthogonality. The data were re-analyzed by using ten decimal places for the α value. When the analysis was re-run with α value set at 1.414213562, the Type I and Type II sums of squares were identical. Hence, it is recommended that more significant digits be used for the coded value to minimize the effect of round-off error.

β Estimates Significance

Once the significant predictors have been identified by conducting the regression analysis using coded values, regression equations are often re-calculated using real-world scores. The advantage of using the real-world values as the independent variable is that meaningful values can be substituted into the real-world valued regression

Table 15. Sums of Squares Summary for Subject Search Time

Variable	df	Estimate (β)	Std. Error	T ratio	Prob
Intercept	1	0.133	0.092	1.444	0.1577
Speech Rate(S)	1	0.088	0.034	2.565	0.0148
Input time-Out (I)	1	0.028	0.034	0.806	0.4255
Music Level (M)	1	0.077	0.034	2.248	0.0310
Age (A)	1	0.085	0.034	2.479	0.0181
S^2	1	0.010	0.054	0.186	0.8534
I^2	1	0.002	0.054	0.031	0.9756
M^2	1	0.004	0.054	0.080	0.9366
A^2	1	0.067	0.054	1.236	0.2246
SI	1	0.029	0.038	0.753	0.4565
SM	1	0.045	0.038	1.169	0.2503
SA	1	0.053	0.038	1.377	0.1774
IM	1	0.091	0.038	2.359	0.0240
IA	1	0.021	0.038	0.555	0.5822
MA	1	0.025	0.038	0.655	0.5170

Variable	df	Type I SS	Type II SS
Intercept	1	2.00052076	0.09878349
Speech Rate(S)	1	0.31151963	0.31151963
Input Time-Out (I)	1	0.03079540	0.03079540
Music Level (M)	1	0.23921560	0.23921560
Age (A)	1	0.29108994	0.29108994
S^2	1	**0.001642240**	**0.001636504**
I^2	1	**0.000045061**	**0.000044137**
M^2	1	**0.000305054**	**0.000302788**
A^2	1	0.07236492	0.07236492
SI	1	0.02684582	0.02684582
SM	1	0.06472422	0.06472422
SA	1	0.08972965	0.08972965
IM	1	0.26356977	0.26356977
IA	1	0.01460361	0.01460361
MA	1	0.02028663	0.02028663

equation (e.g., 240 words per minute instead of +1.414 coded level), and the resultant dependent variable can be readily observed and interpreted. The danger of conducting the analysis using real world values is that the significance of the β weights changes drastically. In Table 16, the regression summary table of the coded values described four variables which have β weights that differed from 0 ($p<0.05$) using the t-statistics. The t-test was conducted under the hypothesis that β did not differ from 0 numerically (H_0: $\beta=0$). If the β weight is different from 0, it is an indication that the particular regressor has a significant amount of predictability in the regression model. The four variables that possessed significant t-ratios were speech rate ($p<0.0148$), age ($p<0.0181$), music level ($p<0.031$), and the interaction of input time-out by music level ($p<0.0239$). Table 17 summarized the same analysis by entering raw scores as the independent variables. Instead of using coded values such as -1.414, -1, 0, 1, and +1.414, the analysis was conducted by utilizing real world values. For example, the five levels of speech rate were entered as 120, 138, 180, 222, and 240 wpm. The entire model now has only one β weight (input time-out by music level) that was significantly different from 0 ($p<0.0238$). Therefore, by conducting the regression equation using raw scores, it is possible to overlook some important predictors which were deemed significant in the analysis using coded values.

Williges (1976) described certain disadvantages of using orthogonal central-composite designs. First, orthogonality will be lost if regression equations are calculated by using real world values. Second, β coefficients will only be orthogonal for second-order models. As described above, the probability of rejecting H_0 changes if real-world values are used; therefore, it is recommended that the significance of the β weights be only interpreted in analysis conducted with coded values, and regression equations calculated using coded values.

Table 16. Regression Equations for Subject Search Time using Raw Scores

Variable	df	Estimate (β)	Std. Error	T ratio	Prob
Intercept	1	2.804	2.001	1.401	0.1701
Speech Rate(S)	1	-0.010	0.012	-0.782	0.4397
Input Time-Out (I)	1	-0.199	0.119	-1.673	0.1032
Music Level (M)	1	-0.039	0.056	-0.701	0.4878
Age (A)	1	-0.038	0.023	-1.686	0.1008
S^2	1	0.000	0.000	0.183	0.8557
I^2	1	0.000	0.006	0.014	0.9886
M^2	1	0.000	0.001	0.084	0.9339
A^2	1	0.000	0.000	1.220	0.2305
SI	1	0.000	0.000	0.752	0.4569
SM	1	0.000	0.000	1.168	0.2506
SA	1	0.000	0.000	1.375	0.1777
I M	**1**	**0.003**	**0.001**	**2.357**	**0.0241**
IA	1	0.000	0.000	0.555	0.5825
MA	1	0.000	0.000	0.654	0.5174

Table 17. Regression Equations for Subject Search Time

Variable	df	Estimate (β)	Std. Error	T ratio	Prob
Intercept	1	0.133	0.092	1.444	0.1577
Speech (S)	1	**0.088**	**0.034**	**2.565**	**0.0148**
Input (I)	1	0.028	0.034	0.806	0.4255
Music Level (M)	1	**0.077**	**0.034**	**2.248**	**0.0310**
Age (A)	1	**0.085**	**0.034**	**2.479**	**0.0181**
S^2	1	0.010	0.054	0.186	0.8534
I^2	1	0.002	0.054	0.031	0.9756
M^2	1	0.004	0.054	0.080	0.9366
A^2	1	0.067	0.054	1.236	0.2246
SI	1	0.029	0.038	0.753	0.4565
SM	1	0.045	0.038	1.169	0.2503
SA	1	0.053	0.038	1.377	0.1774
IM	1	**0.091**	**0.038**	**2.359**	**0.0240**
IA	1	0.021	0.038	0.555	0.5822
MA	1	0.025	0.038	0.655	0.5170

Optimal Settings of Variables

The purpose of studying different variables using response surfaces was to predict optimal combinations of variables. This study used three major groups of dependent measures: search time, keypresses, transcription errors. Therefore each combination of variables was unique for each dependent measure.

As presented in the results section, each dependent measure has its own unique optimal solutions. Although all of the calculated optimal points are saddle points, canonical analysis should be conducted to determine if these points represent true optimal values. However, most of the calculated optimum values lie within the range of variables studied, the optimal values do have some merit. Therefore, a heuristic procedure was used in selecting the optimal combination.

First, optimal speech rates were calculated as 266.21 wpm, 154.84 wpm, and 175.76 wpm for the measures of search time, transcription errors, and extra keypresses. However, the 266.21 wpm was calculated using the dependent variable subject search time. The subject search time variable did not take into the account of transcription errors. From the regression equation, a speech rate of 266.21 would have resulted in a transcription error rate of close to 37 percent (1.48 words/minute). Therefore the 154.84 wpm is selected because both transcription error and extra keypresses decreased with decreasing speech rate.

Second, input time-out was found to have the value of 1.76 seconds, 5.89 seconds, and 3.37 seconds for the dependent measures of subject search time, transcription errors, and extra keypresses respectively. Since input time-out was in no way affecting transcription average, the value of 5.89 seconds can be discarded. Input time-out was a significant predictor for both total and subject search time. From Figure 5, there is quite a clear evidence that search time decreased when input time-out was

decreased. Therefore, 1.76 seconds seemed to be a logical choice because the value 3.37 seconds only adds to subject and system search time.

Third, music level was found to have the optimal value of 44.06 dB(A), 42.75 dB(A), and 57.23 dB(A) for the dependent measures of subject search time, transcription errors, and extra keypresses respectively. Music level greater than 50 dB(A) has a deleterious effect on both transcription error, subject search time, total search time, and extra keypresses. The lower of the two music level of 42.75 dB(A) is recommended.

The last variable is age. Interestingly, the three ages are 20.52, 23.44, and 14.04 for the dependent measures of subject search time, transcription errors, and extra keypresses respectively. It is difficult to predict what age of the users will perform best on this system given age is not the only characteristic that affects the performance of using a synthetic-speech computer system. However, all three ages seemed to be quite a logical selection; the younger population certainly performed better than the older subjects.

CONCLUSIONS

For the telephone information system, the results of this study can be summarized as follows:

- For information retrieval tasks, when the search time is not affected greatly by the rate of the speech synthesizer (i.e. listening to short menu options), speech rate should be set in between 120 - 150 words per minute. However, slow speech rate such as 120 words per minute would lead to slower search time and it would increase search time if long information messages are used.

- Input time-out can be set as short as two seconds because subject search time was not affected by the changes in input time-out values.

- Noise level is best under 50 dB(A). Subjects' ability to understand synthetic speech decreased rapidly when noise level was above 50 dB(A); however, better tolerance of noise might be achieved by using telephone handsets instead of speaker telephones. Loud noise level (> 60 dB(A)) hampered subjects' ability to perceive synthetic speech and was reported as annoying to most subjects.

- Older subjects performed less well in most of the tasks while subjects as young as 14 had very little trouble in using the system.

- Although the subjects had very little experience with synthetic speech, they were able to detect that the rate of 222 words per minute was faster than the rate of 138 words per minute.

- When users are trying to obtain information from a speech system, contextual information should be placed near the end of a message to lessen the burden of users' short-term memory.

For methodological issues, the results of this study lead to the following recommendations:

- When performing regression analysis using central-composite designs, the analysis should be conducted using a large number of decimal places for the α values.

- Use coded values instead of real world-values to ensure the orthogonality of the variances of each regressor.

- To determine the significance of the regressor, do not rely on just the t-statistics for testing the predictability of the β weights. Conduct analysis of variance using the coded values to determine the significant predictors.

- The optimal values calculated for age were 14 to 20 years. This finding does not contribute as much information as the other three variables because usually a designer is not interested in selecting a optimal age group but rather to allow the system to be used in a wide age range without compromising the system performance. However, if user's age of the system is known, the prediction equation can be very useful in obtaining information of the system performance by setting the age as a constant and varying the remaining variables.

Recommendations for further research:

- Since young subjects performed quite well in this study, even younger (< 14 years old) subjects might be recruited to see at what age the younger population will have problems interacting with telephone inquiry systems.

- The use of input time-out values of less than two seconds with finer increments than one second might be worth pursuing because it is clear that in this experiment, the short input time values of two or three seconds had no noticeable effects on user behavior.

- To investigate the effects of speech rate, use longer menu options to determine if speech rate can significantly affect search time.

REFERENCES

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.

Aucella, F. A., and Ehrlich S. F. (1986). Voice messaging enhancing the user interface based on field performance. In *Proceedings Chi'86, Human Factors in Computer Systems.* (pp. 156-161). Boston, ACM, New York.

Beaudet, D. B. (1988). *The effects of 16 variables on a telephone information system which uses synthetic speech.* Unpublished masters thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Beranek, L. L., Blazier, W. E., and Figwer, J. J. (1971). Preferred noise criterion (PNC) curves and their application to rooms. *Journal of the Acoustical Society of America, 50,* 1223-1228.

Box, G. E. P., and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society, Series B (Methodological),* 13, 1-45.

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

Gould, J. D., and Boise, S. J. (1984). Speech filing - An office system for principals. *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.

Halstead-Nussloch, R. (1989). The design of phone-based interfaces for consumers. In *Proceedings Chi'89, Human Factors in Computer Systems.* (pp. 347-352). Austin, ACM, New York.

Herlong, D. W. (1988). *Effects of voice coding and speech rate on a synthetic speech display in a telephone information system.* Unpublished masters thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

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.

Kryter, K. D. (1972). Speech communication. In H. P. Van Cott and R. C. Kinkade (Eds.), *Human engineering guide to equipment design*(pp. 161-226). New York: John Wiley & Sons.

Lee, F. F. (1983). Time compression and expansion of speech by the sample method. In J. S. Lim (Ed.), *Speech Enhancement* (pp.286-290). London: Prentice Hall.

Maurer, J. F., and Rupp, R. R. (1979). *Hearing & Aging.* New York: Grune & Stratton.

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.

Pisoni, D. B. (1979). *Some measures of intelligibility and comprehension.* In Research on Speech and Perception Report No. 5 (pp. 3-47). Bloomington, IN: Indiana University.

Roberts, T. L., and Engelbech G. (1989). The effects of device technology on the usability of advanced telephone functions. In *Proceedings Chi'89, Human Factors in Computer Systems.* (pp. 331-337). Austin, ACM, New York.

Rosson, M. B., and Mellen, N. M. (1985). *Behavioral issues in speech-based remote information retrieval* (Research Report RC11028 [49528]). 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. (1985). *SAS User's Guide: Statistics, Version 5 Edition.* 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., and Castellan, N. J. (1988). *Nonparametric statistics for the behavioral Sciences.* New York: McGraw-Hill.

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.

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.

Spoor, A. (1973). Presbyacusis in relation to noise-induced hearing loss. In *Proceedings of the International Congress on Noise as a Public Health Problem.* (pp. 257-266). The U. S. Environmental Protection Agency.

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.

Wichansky, A. M. (1987). Learning and using office automation on personal computers: a voice/phone application. In *Proceedings of the Human Factors Society 31th Annual Meeting* (pp. 266-269). Santa Monica, CA: Human Factors Society.

Williams, K. (1978). An introduction to the assessment and measurement of sound. In D. M. Lipscomb and A. C. Taylor (Eds.), *Noise control: handbook of principles and practices* (pp. 33-60). New York: Van Nostrand Reinhold.

Williges, R. C. (1976). Research note: modified orthogonal central-composite designs. *Human Factors, 18,* 95-98.

Williges, R. C. (1981). Development and use of research methodologies for complex system simulation experimentation. In M. J. Morall and K. F. Kraiss (Eds.), *Manned system design, methods, equipments and applications* (pp. 59-87). New York: Plenum.

Williges, R. C., and North R. A. (1973). Prediction and cross-validation of video cartographic symbol location performance. *Human Factors, 15,* 321-336.

Williges, R. C., and Williges, B. H. (1982). Modeling the human operator in computer-based data entry. *Human Factors,* 24, 285-299.

Witten, I. H., and Madams, 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

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)

Appendix II

Participant's Informed Consent Form

The following experiment is a study concerning the evaluation of a telephone based information system. The experimenter will conduct a hearing test before the experiment is to take place. The purpose of the hearing test is to determine whether your hearing meets the criteria we have established for participating in our experiment. This is **NOT** a professional hearing test and the results should not be considered as an accurate description of your hearing. If your hearing meets the criteria that we have established, you will be asked to participate in our experiment. You will be required for one session which will last approximately two hours. You will be paid at a rate of $5/hour.

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 Jimmy K. Wu. All of these people can be contacted at the following address and phone number:

Human-Computer Interaction Laboratory
530 Whittemore Hall
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061
(703) 231-4602

Further comments or questions can be addressed to Ernest Stout, 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. Ernest Stout
Office of Sponsored Research Programs
301 Burruss Hall
Virginia Polytechnic Institute and State University
(703) 231-5281

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

Parent or Guardian's Signature
(If Participant is under 18 years of age)

_____

Witness' Signature

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

_____

_____

_____

_____

Appendix III

Subject's 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.

Appendix IV

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 uncertain ... very 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 difficult ... very 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 difficult ... very easy

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

Appendix V

Subjective Rating Scales

Subjective ratings collected after all the target searches are completed.

Ease-of-use

1-----2-----3-----4-----5-----6-----7
very difficult ... very 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 unintelligible ... very 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 unnatural ... very 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 slow — very fast

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

Input Timeout

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

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

Background Music

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

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

Appendix VI
Summary Information

Assessing Human Performance Trade-Offs of a Telephone-Based Information System

Little research effort has been devoted to human interaction with telephone information systems. A study investigating the effects of synthesized speech rate, time available for user input, subject age, and background music level on human behavior in an interactive telephone-based information system was conduted. Subjects searched a fictitious department store database for 16 specific store items and transcribed 16 information messages which were spoken by a computer speech synthesizer.

Paritcipants rated certain features of the system and performance measures were also collected from the subjects on an on-line basis. Performance was evaluated by search time, and number of errors in the transcription tasks. Two seconds was found to be an optimal time for users to enter their selection. The computer synthesized speech rate should be set close to 150 words per minute. Background music or noise level should be kept below 50 dB(A); sound level above 50 dB(A) seriously affected user's ability to understand synthetic speech. Most of the younger subjects (age 14 - 22) performed better in this study than the older subjects (age 36-62).

The vita has been removed from the scanned document