UNCLASSIFIED

DTIC FILE COPY

(1)

AD-A196 295

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFIT/CI/NR 88- 7	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) EFFECTS OF VOICE CODING AND SPEECH RATE ON A SYNTHETIC SPEECH DISPLAY IN A TELEPHONE INFORMATION SYSTEM		5. TYPE OF REPORT & PERIOD COVERED MS THESIS
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) DAVID W. HERLONG		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS AFIT STUDENT AT: VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE 1988
		13. NUMBER OF PAGES 127
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) AFIT/NR Wright-Patterson AFB OH 45433-6583		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

DISTRIBUTED UNLIMITED: APPROVED FOR PUBLIC RELEASE

DTIC ELECTE AUG 0 3 1988 S D

D

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

SAME AS REPORT

18. SUPPLEMENTARY NOTES

Approved for Public Release: IAW AFR 190-1
LYNN E. WOLAVER Lynn Wolaver 18 July 88
Dean for Research and Professional Development
Air Force Institute of Technology
Wright-Patterson AFB OH 45433-6583

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

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

ATTACHED

88

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Effects of Voice Coding and Speech Rate

On A Synthetic Speech Display

In A Telephone Information System

by

David W. Herlong

Major, United States Air Force

127 pages

Master of Science

in

Industrial Engineering and Operations Research

(ABSTRACT)

Despite the lack of formal guidelines, synthetic speech displays are used in a growing variety of applications. Telephone information systems permitting human-computer interaction from remote locations are an especially popular implementation of computer-generated speech. Currently, human factors research is needed to specify design characteristics providing usable telephone information systems as defined by task performance and user ratings. Previous research used nonintegrated tasks such as transcription of phonetic syllables, words, or sentences to assess task performance or user preference differences. This study used a computer-driven telephone information system as a real-time, human-computer interface to simulate applications where synthetic speech is used to access data. Subjects used a telephone keypad to navigate through an automated, department-store database to locate and transcribe specific information messages. Because speech provides a sequential and transient information display, users may have difficulty navigating through auditory databases. One issue investigated in this study was whether use of alternating male and female voices to code different levels in the database hierarchy would improve user search performance. Other issues investigated were basic intelligibility of these male and female voices as influenced by different levels of speech rate. All factors were assessed as functions of search or transcription task performance and user preference. Analysis of transcription accuracy, search efficiency and time, and subjective ratings revealed an overall significant effect of speech rate on all three groups of measures but no significant effects for voice type or coding scheme. Results were used to recommend design guidelines for developing speech displays for telephone information systems.

Effects of Voice Coding and Speech Rate

On A Synthetic Speech Display

In A Telephone Information System

David W. Herlong

May 1988

Effects of Voice Coding and Speech Rate

On A Synthetic Speech Display

In A Telephone Information System

by

David W. Herlong

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

Robert C. Williges, Chairman

Dennis L. Price

Dennis L. Price

Beverly H. Williges

Beverly H. Williges

May 12, 1988

Blacksburg, Virginia

Effects of Voice Coding and Speech Rate

On A Synthetic Speech Display

In A Telephone Information System

by

David W. Herlong

Robert C. Williges, Chairman

Industrial Engineering and Operations Research

(ABSTRACT)

Despite the lack of formal guidelines, synthetic speech displays are used in a growing variety of applications. Telephone information systems permitting human-computer interaction from remote locations are an especially popular implementation of computer-generated speech. Currently, human factors research is needed to specify design characteristics providing usable telephone information systems as defined by task performance and user ratings. Previous research used nonintegrated tasks such as transcription of phonetic syllables, words, or sentences to assess task performance or user preference differences. This study used a computer-driven telephone information system as a real-time, human-computer interface to simulate applications where synthetic speech is used to access data. Subjects used a telephone keypad to navigate through an automated, department-store database to locate and transcribe specific information messages. Because speech provides a sequential and transient information display, users may have difficulty navigating through auditory databases. One issue investigated in this study was whether use of alternating male and female voices to code different levels in the database hierarchy would improve user search performance. Other issues investigated were basic intelligibility of these male and female voices as influenced by different levels of speech rate. All factors were assessed as functions of search or transcription task performance and user preference. Analysis of transcription accuracy, search efficiency and time, and subjective ratings revealed an overall significant effect of speech rate on all groups of measures but no significant effects for voice type or coding scheme. Results were used to recommend design guidelines for developing speech displays for telephone information systems.

Dedication

To God for my intelligence and my mother for the desire to use it.

Acknowledgements

Support for this research was provided by the National Science Foundation under direction of Dr. H. E. Bamford, Jr.. Special recognition is due Dr. R. C. Williges and B. H. Williges, co-principal investigators of the grant study of which this research was a part. Their advice and counsel became a valuable (and valued) part of this effort. I also wish to thank Dr. Dennis L. Price for providing me with thoughtful suggestions which proved essential to research design and conduct. I offer my thanks and grateful appreciation to Dr. Robert D. Dryden for substituting during the final defense. Without Calvin L. Selig, Systems Manager for the Virginia Tech Human Computer Interaction Laboratory and software author for this project, sophisticated simulation of a telephone information system would not have been possible. Freely given assistance and constructive criticism from Douglas B. Beaudet and Peter J. Merkle, Jr., office-mates and co-workers, greatly enhanced the quality of my study.

Finally, I especially wish to thank my loving wife, Elizabeth V. Herlong, for her confidence in me and enthusiastic encouragement which sustained me throughout the program.

Table of Contents

List of Illustrations

List of Tables

Introduction

Overview

Modern speech research involving electronic analysis of speech began with the introduction of the sound spectrograph developed by the Bell Telephone Laboratories in 1946 and Franklin Cooper's "pattern playback" machine constructed in 1950 at the Haskins Laboratories (Pisoni, 1982). Synthetic speech research remained the province of large research centers until the late 1970's. According to Bristow (1984), the innovation of Very Large Scale Integration (VLSI) devices in 1977 initiated a "[synthetic] speech revolution". Reliable performance and attractive cost of VLSI's resulted in a marked increase of synthetic speech research and rapid introduction of synthetic speech displays to the public domain. Figure 1 on page 2 depicts a summary of the history of synthetic speech concept and hardware development (see Appendix A for references used in Figure 1).

Commercial developers of speech synthesizers did not wait for further research. Instead, synthetic speech displays were implemented in absence of empir-

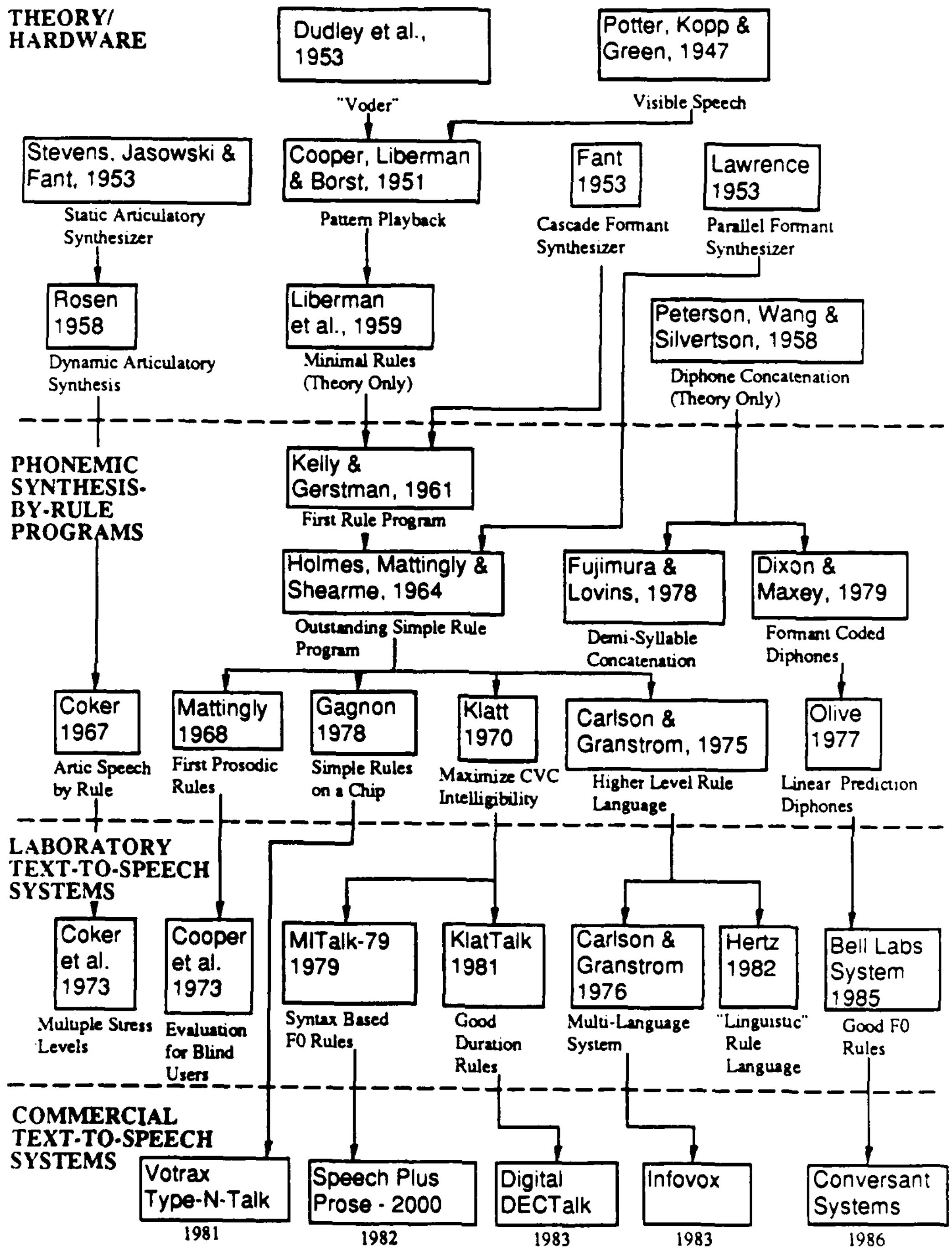

Figure 1. Research Summary of Synthetic Speech Concepts and Hardware (From Klatt, 1986)

ically derived guidelines (Pisoni, 1982). This parallel progression of research and operational implementation continues today. And it is text-to-speech synthesizers that promise the greatest utility for applications in which unrestricted English text must be converted into speech such as information retrieval by phone (Slowiazcek and Nusbaum, 1985; Allen, 1981). Current speech technology gives us the opportunity to make the telephone a terminal, thereby taking greater advantage of a device one author termed the "most powerful communications tool in human history" (McHugh, 1986). Telephones have a large user population allowing access to telephone-based information systems from practically anywhere. Additionally, those who might be otherwise intimidated by computers may more freely accept using a familiar and simple device such as the telephone as a terminal for computer-driven information systems (Labrador and Pai, 1984). Yet, we have few if any guidelines for using synthetic speech displays in telephone information systems.

Purpose

This study addressed lack of guidelines for telephone-based information systems by investigating effects of voice type and speech rate on task performance of a synthetic speech display. Measures of intelligibility and search efficiency were used to detect performance differences and subjective ratings to assess user preferences and impressions. A major question of this study was whether alternating male and female synthetic voices as an informational coding scheme improved performance in an automated database as compared to using a single voice to present all information. Related to this issue was whether one voice was more intelligible than the other for pronouncing key-words and sentences. Finally, this study examined the effect on task performance and user preferences of increasing speech rate beyond optimum rates demonstrated in previous research. Previous study results suggest a performance optimum of 180 wpm for DECtalk's Perfect Paul voice (Merva and Williges, 1987). This study continued the inquiry of optimum rate by comparing Perfect Paul to DECtalk's Beautiful Betty voice, also found highly intelligible in earlier research (Greene, Manous, and Pisoni, 1984).

Literature Review

Methods of Speech Synthesis

For this study, the term *synthetic speech* refers to speech generated entirely by rule or algorithms without the aid of an original, human recording (Simpson, McCauley, Roland, Ruth, and Williges, 1985). Computers also use other methods of speech generation such as *digitized speech* and *analysis-synthesis*. These alternate methods of producing synthetic speech may feature better voice quality than speech synthesized by rule but suffer disadvantages not shared by rule-generated speech.

Digitized Speech

Speech synthesis by rule differs from digitized speech which is human speech recorded digitally and then (usually) transformed into a more compressed data format. Digit:l recording processes may sample human speech up to 8000 or more times per second. Fidelity to the original signal and hence, intelligibility, is excellent at such rates but massive amounts of storage capability are required to store the digitized information (Sanders and McCormick, 1987). Storage limitations lead to fixed-sized vocabularies which must also be updated to add new words. Furthermore, since digitized speech depends on an original source, voice variety is fixed for a recording. To use additional voices in a digitized speech display compounds storage problems mentioned earlier. The unlimited variety of human voices available for a digitized speech display also imparts unique problems of variability in its research (Simpson, *et al.*, 1985). Research replication using digitized speed would require either the same voice or one similar as selected by standard voice parameters. Additionally, guideline standardization becomes very difficult with a virtually unlimited variety of human voices for digital recording sources.

Synthesis by Analysis

Analysis-synthesis methods electronically model the human voice mechanism to produce speech sounds (Sanders and McCormick, 1987). The source speech wave is analyzed along certain parameters which are encoded by the speech analyzer and stored. This method, also known as *waveform sampling*, differs from digitized speech which encodes the actual speech wave and requires far more computer memory to store speech information than does speech synthesized by rule. For example, analysis-synthesis using a common analog-to-digital conversion requires about 64,000 bits per second for uncompressed speech (8000 samples per second to capture up to 4000 Hertz (Hz), multiplied by 8 bits per sample) (Kaplan and Lerner,1985). The same, approximate memory requirements used by digitized speech result in very natural (human-like) speech. However, speech produced by analysis-synthesis tends to sound awkward and unnatural because of a lack of *coarticulation* or the natural blending and modification of speech sounds caused by words and phonemes that precede and follow a particular sound. A *phoneme* can be thought of as the smallest speech sound that can change the meaning of a word, but the phoneme is really more a theoretical definition than a precise definition of the spoken segments of our speech alphabet (Kantowitz and Sorkin, 1983). Some (Simpson, *et al.*, 1985; Flanagan, 1972) refer to analysis-synthesis methods as digitized speech since it uses a digital data-compression technique.

Speech Synthesis by Rule

Speech generated by rule uses stored dictionaries of elementary speech segments and sets of rules for combining them and for stressing particular sounds or words that produce the *prosody* of speech (Sanders and McCormick, 1987). Prosody is the rhythm or singsonq quality of natural speech. Unlike digitized speech, rule-generated or synthetic speech requires far less computer memory since it makes direct translation of text into speech. As an example, *formant (resonant frequency) synthesis*, one of two methods used to synthesize speech by rule, requires a data rate of 100 bits per second based on a typical rate of 12 phonemes per second with each phoneme characterized by an 8-bit code (Kaplan and Lerner, 1985). This memory requirement is far less than the 64,000 bits per second required by analysis-synthesis or digitized speech methods. Formant synthesis simulates the formants or resonances of the vocal tract and is used by Digital Equipment Corporation's, DECtalk, the speech synthesizer used in this study. *Linear predictive coding*

(LPC), the other rule-generated, synthetic speech method uses a mathematical representation of the vocal tract as acoustic tubes to produce speech.

Another advantage of rule-generated, synthetic speech possessed by neither digitized nor analysis-synthesis speech is direct, text translation which provides another name for this method, *text-to-speech*. Rule-based speech synthesizers also feature several file or default voice types making standardization of research and resulting guidelines more practical. Consequently, synthetic speech systems do not depend on human speakers for new vocabularies as do digitized speech or analysis-synthesis speech which must use the same human speaker in order to sound consistent (Simpson, *et al.*, 1985). However, the best synthetic speech has yet to achieve a voice quality comparable to the best of other methods. This limitation has made intelligibility the prime variable of interest in most synthetic speech research with many related issues still unresolved.

Perception of Synthetic Speech

With few exceptions, previous research has consistently demonstrated synthetic speech to be less intelligible than natural, human speech except under optimum conditions of low noise and high context (Pisoni and Hunnicut, 1980; Greene, *et al.*, 1984). This lower intelligibility produces two effects: either the information presented by synthetic speech is not heard or remembered accurately, or the additional effort required to understand it interferes with other tasks being carried out at the same time (Cooper, 1987). Less clear are reasons behind the lower intelligibility. However, researchers usually consider problems of synthetic speech intelligibility to lie in human processes of speech perception and information processing. Luce, Feustel, and Pisoni (1983) have suggested comprehension of synthetic speech places a greater cognitive load on the listener because synthetic speech does not possess cues present in natural, human speech. Additionally, Nusbaum, Dedina and Pisoni, (1984) postulate a possible increase in short term memory requirements. Models of human information processing are necessary to consider problems of synthetic speech intelligibility in the context of short term memory.

Information Processing Theory

Broadbent (1958) formulated the limited-capacity channel model which has proved to be a milestone in human information processing research (Kantowitz and Sorkin, 1983). As depicted in Figure 2 on page 10, this formulation was characterized by four features:

- The whole nervous system is regarded as a single channel, having a limit to the rate at which it can transmit information.
- The limited-capacity portion of the nervous system is preceded and protected by a selective filter.
- This "filter" is preceded by a buffer or temporary (short-term) store which could hold any excess information arriving by channels other than the one selected.
- A long-term store kept information passing through the limited-capacity system in the form of a record of the conditional probability that events of one kind are followed by events of another kind.

This "reasonable first approximation of human capabilities in most tasks" has since been modified by Broadbent (1971, 1982) and challenged by some (Kantowitz, 1974; Kinsbourne, 1981; Lane, 1981).

Most challenges to Broadbent's model reveal the bottleneck in information processing as represented by the limited-capacity channel is not as straightforward

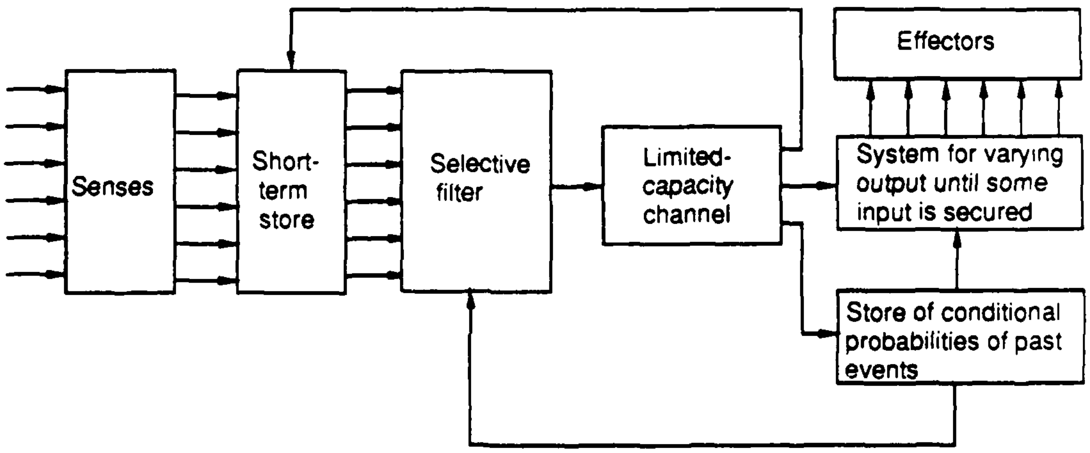

Figure 2. Original limited-capacity channel model (From Broadbent, 1958)

in practice as originally thought. The basic tenet of the limited-capacity model is that humans can transmit information only at a finite rate with output of one stage feeding directly into the next stage. — a *serial processing* function. More recent models emphasize *hybrid processing*. which use both serial and *parallel processing* in the same activity. Parallel processing occurs when several stages simultaneously have access to the same output of another stage (McCormick and Sanders, 1982). Unlike the limited-capacity model, hybrid models allow information to enter in parallel with no bottleneck (Kantowitz and Sorkin, 1983). Bottlenecks occur only when responses must be emitted.

Although hybrid models are still undergoing revisions and challenges characteristic of empirical methodology, most information theorists agree to existence of short term memory (STM) — a function critical in synthetic speech perception. Research efforts of Atkinson and Shiffrin (1968) suggest STM acts not only as a repository for new information but incorporates a working memory responsible for decision making, problem solving, and the general flow of information within the memory system. *Rehearsal*, the overt or covert repetition of information, is one of the control processes used to govern functions within the STM's working memory by maintaining information within STM. Miller's classic paper, "The Magical Number Seven Plus or Minus Two" reported research that demonstrated people can remember approximately seven items (Miller, 1956). More items could be recalled if combined into meaningful "chunks", but the number of chunks (not bits) remained approximately seven. Miller's view is still held to be generally correct with further research demonstrating memory capacity to be influenced also by such factors as acoustic similarity and word length (Conrad and Hull, 1964; Baddeley, Thomson, and Buchanan, 1975).

Research findings stemming from these theories hold several implications for designers of human-computer interfaces which use synthetic speech as a display. In a series of three experiments, Luce, *et al.* (1983) compared subjects' recall for synthetic and natural lists of monosyllabic words using the MITalk speech synthesizer. From their results, they concluded difficulties in perception and comprehension of synthetic speech are due in part to increased processing demands in short-term memory (STM). A subsequent study by Nusbaum, *et al.* (1984) investigated two opposing hypotheses for these increased processing demands imposed on STM. The first hypothesis held synthetic speech to be simply equivalent to "noisy" natural speech. That is, basic cues of synthetic speech were obscured, masked or physically degraded in a way similar to that of natural speech in noise. A second, counter hypothesis postulated synthetic speech to be perceptually impoverished relative to natural speech both in degree and kind. Using three speech synthesizers and recordings of natural voice in four levels of noise, Nusbaum, *et al.* had 83 undergraduates listen to one of these seven speech sources speak 48 consonant-vowel (CV) syllables. Distribution patterns of errors and confusions by subjects clearly supported the hypothesis that synthetic speech is not perceived like natural speech but is some sense, impoverished.

Synthetic Speech Dependent Variables

Performance Measures

Synthetic speech research uses performance and preference measures to assess independent variable effects on perception as reflected by dependent variable constructs. Intelligibility, the fundamental dependent variable construct, is defined operationally as the percentage of speech units correctly recognized by a human listener out of a set of speech hits such as words, sentences, phonemes or the perceptual acoustical features of those phonemes (Simpson *et al.*, 1985). Performance measures of intelligibility for synthetic speech research were borrowed from traditional communications research and include: Modified Rhyme Test (MRT), both open and closed set (Fairbanks, 1958; House, Williams, Hecker, and Kryter, 1965); Harvard Psycho-Acoustic Sentences (Egan, 1948); and Haskins Semantically Anomalous Sentences (Nye and Gaitenby, 1974). Standardization, a strong advantage of these measures, allows researchers to compare results across different conditions such as performance of different speech synthesizers or different researchers to compare study findings. However, there has been recent criticism of these measures and the MRT in particular.

O'Malley and Caisse (1987) point out the original MRT was never intended to be a measure of human speakers' ability to *produce* intelligible speech but developed instead to measure transmission, not several, serious deficiencies:

- MRT results are more unstable with computer speech than with human speech because of a strong learning curve (training effect) associated with listening to synthetic speech.

- The MRT sound list is too limited, testing only 300 monosyllables thus ignoring vowel phonemes, some consonants and all consonant clusters.
- The MRT only tests isolated words and does not reflect that in computer speech, consonants occur next to silence less than 5% of the time. Except for menus as used in this study, most speech occurs in sentences (also used in this study), and putting words together is the most difficult task for phoneme-to-speech modules.
- Few MRT tests reported so far have been conducted in a telephone environment with its accompanying noise and bandwidth limitations thus ignoring telephone involvement in 90% of computer speech applications.
- Vendors attempt to tune their systems to the 300 words found in the MRT.

Sentences also have their advantages and disadvantages when used in intelligibility studies. Sentences are more appropriate for research purposes when used for evaluating telephone information systems in which sentences are the usual unit of information of interest. However, considerable differences in systems must exist before significant differences will be obtained in transcription scores. Psychological factors (meaning, context, rhythm) make sentence test scores difficult to analyze and interpret. For extensive testing, a large number of sentences is required since the listener will remember sentences. Furthermore, sentences used in actual, auditory displays tend to be unique both in vernacular and context because of the particular, application setting. Consequently, researchers must employ systematic sentence construction techniques in order to generalize results and attempt derivation of global principles of sentence usage in synthetic speech displays.

Preference Measures

Preference measures have been either inferred from performance data or directly measured using self-report measures such as subjective ratings and comparisons. Listener impressions of naturalness, pleasantness, and acceptability as compared to a human voice are the usual dimensions polled. Other dimensions such as confidence and appropriateness are among many variations devised by researchers. Rating-scale types have included Likert scales (one to seven numerical ratings), descriptively anchored scales ("very human" as opposed to "very machine-like"), and bipolar scales ("harsh" versus "soothing"). Open-set queries have no researcher-provided response to choose from and though the data is less quantifiable, it often proves invaluable to the researcher/designer. Because of their non-parametric qualities, subjective rating methods are difficult to analyze with parametric statistics. There have been attempts to relate subjective ratings to objective measures of speech intelligibility (Barnwell, 1982; Voiers, 1977) and thus impart parametric attributes.

One such measure is the Diagnostic Rhyme Test (DRT) described by Voiers (1983). Subjects compare relative intelligibility of 96 rhyming word pairs that differ by a single acoustic feature or attribute in the initial consonant. The six attributes are: voicing, nasality, sustention, sibilation, graveness, and compactness. Widely used within the Department of Defense (DOD), the DRT has the advantage of providing highly reliable and repeatable scores that can be used to make comparisons even among systems evaluated at different times (Schmidt-Nielsen, 1985). However, potential users of voice systems dislike the DRT because they lack a reference frame by which to evaluate DRT scores. Instead, they prefer "realistic" tests despite the fact that such tests are often unrepeatable because results are confounded by such irrelevant variables as noise, distractions, and interruptions (Schmidt-Nielsen, 1985).

Pratt (1987) used another subjective or preference measure, Multi-Dimensional Scaling (MDS), in which subjects rate the dissimilarity between members of a set of stimuli. In this measure, subjects are presented with pairs of stimuli and instructed to assign a numerical value to the degree of dissimilarity between members of each pair. Data reduction techniques produce estimates of dissimilarity which the experimenter is then required to interpret intuitively. Yet another preference measure is the Semantic Differential Scaling (SDS) developed by Osgood, Suci, and Tannenbaum (1957). In the SDS, subjects rate stimuli by selecting a point on a numbered scale which has been anchored at either end with antonymous adjectives. This method is very similar to the bipolar, seven-point scales used in this study.

Selected Independent Variables

Voice Type

Early speech synthesizers had one voice unique to the machine. Now, synthesizers are capable of producing an almost endless variety of voices by manipulating adjustable parameters. The DECtalk version 2.0 used in this study allows experimenter control over 32 different parameters as well as possessing 9 default voices. Consequently, intelligibility of different synthesizer voices as compared to each other has been a natural, research focus. Some speech synthesizers have achieved intelligibility rates of 100% by careful manipulation of parameters and algorithms for certain words. Such a file of "customized words" is called an *exception dictionary*. Indeed, in certain conditions of noise or distractions, some subjects have rated synthetic speech more intelligible than natural speech citing its distinctive qualities (Simpson, 1983; Simpson and Williams, 1980). In a study designed to assess synthetic speech qualities, Rosson and Cecala (1985) manipulated four parameters of head size, pitch, richness and smoothness using sixteen perceptual-scale ratings to derive preference measures. However, research involving methodical manipulation of individual synthetic speech parameters to evaluate performance is still lacking.

Instead, most research has used the default voices of speech synthesizers. Greene, *et al.*, (1984) compared the DECtalk, version 1.8 to earlier evaluations of the Prose-2000, version 8-84; the MITalk-79; and the Type-n-Talk, version 3-82. Using the open- and closed-set Modified Rhyme Test, the Harvard Psycho-Acoustic sentences, and the Haskins Semantically Anomalous Sentences, they found the DECtalk unit the most intelligible with error rates roughly half the size of error rates observed in earlier studies. Of the two default DECtalk voices evaluated, Perfect Paul appeared more intelligible than Beautiful Betty. Paul and Betty are male and female voices respectively, which according to listeners, sound "middle-aged with an occasional accent." A comparison yet to be made and a focus of this study is whether these two most intelligible voices, Paul and Betty, differ significantly in intelligibility for sentences as well as isolated words and word units.

Speech Rate

Early research favored a speech rate of approximately 150 wpm. Simpson and Marchionda-Frost (1984) using a Votrax ML-1 synthesizer investigated three word rates: 123, 156, and 178. Although they found intelligibility unaffected by speech rate, subjects reported a subjective preference for 156 wpm. Lack of a performance effect on intelligibility resulted from Simpson and Marchionda-Frost training their subjects to 100% intelligibility on a small, highly-constrained vocabulary thus maximizing contextual cues (Slowiaczek and Nusbaum, 1985). In a two-study series, Slowiaczek and Nusbaum (1984) investigated the performance effects of 150 wpm and 250 wpm on intelligibility using a Prose-2000 speech synthesizer. Their findings confirmed Simpson and Marchionda-Frost's subject preference for 150 wpm. Waterworth and Lo (1984) in investigating the effects of six rates (63, 82, 103, 121, 130, and 150 wpm), found messages at the higher rates to be more intelligible though no differences were statistically significant. Their study compared natural voice to four synthesizers, three of which were text-to-speech synthesizers: Votrax CDS-II, Prose-2000 and the Microspeech-2.

Recent research findings, however, indicate an optimum speech rate of 180 words per minute (wpm) for synthetic speech, a rate which approximates the average for conversational speech. This optimum was for speech produced by the DECtalk synthesizer's Perfect Paul voice (Merva and Williges, 1986; Merva, 1987). In one study (Merva and Williges, 1986), a rate of 250 wpm was shown to be significantly less intelligible than a 180 wpm rate. In a follow-on study, Merva (1987) compared three speech rates of 150 wpm (the preferred rate reported by Simpson and Marchionda-Frost, 1984), 180 wpm, and 210 wpm and again found performance measures indicating 180 wpm as the optimum rate. Both studies, however, used sentences as the audible targets. Sentences provide more linguistic, contextual clues than single words (Simpson and Williams, 1975), but single words or small phrases are necessary for menu selection choices in auditory databases. Further investigation of relatively, high speech rates may enable increases in auditory display rates allowing users to scan messages more quickly (O'Malley and Caisse, 1987). Also, no study has systematically investigated the possible interaction of voice type and speech rate on intelligibility. This study addressed all those issues.

Information Coding

The issues investigated in this study pertain not only to the intelligibility of synthetic speech but to principles of auditory displays as well. Most human factors research efforts in synthetic speech attempt to refine and expand guidelines for display design and implementation. The starting point for many has been Deatherage's (1972) comparison table for auditory and visual display forms. Though quantitative guidelines derived from research findings are still forthcoming, designers at least can remain aware to problems especially those revealed in information processing studies. As an example, Kidd (1982) provides several problems pertinent to auditory displays:

- A user's short term memory storage capacity is severely limited with any new input decaying rapidly unless constantly rehearsed.
- Any problem solving, decision making or other information processing severely restricts the user's ability to carry out the necessary rehearsal of new information.
- Synthetic speech (currently) requires more effort to process than human speech.
- The user cannot control the rate at which information is received.
- The user is unable to rapidly scan the menu list in search of a target item and instead must hear each item individually.
- Possible user anxiety may result from not knowing how many menu items will have to be remembered during an interaction.

Sanders and McCormick (1987) do offer tentative guidelines for synthetic speech display implementation (see Table 1 on page 20) "gleaned" from these sources: Simpson and Williams, 1980; Thomas, Rosson, and Chodorow, 1984; and Wheale, 1980.

System designers should also attempt to take advantage of chunking while remembering the limited STM capacity by providing clues about the classification

Table 1. Synthetic Speech Implementation Guidelines

1. Voice warnings should be presented in a voice that is qualitatively different from other voices that will be heard in the situation.
2. If synthesized speech is used exclusively for warnings, there should be no alerting tones before the warning.
3. If synthesized speech is used for other types of information in addition to warnings, some means of directing attention to the warning might be required.
4. Maximize intelligibility of the messages.
5. For general-purpose use, maximize user acceptance by making the voice as natural as possible.
6. Consider providing a replay mode in the system so users can replay the message if they desire.
7. If a spelling mode is provided, its quality may need to be better than that used for the rest of the system.
8. Give the user the ability to interrupt the message; this is especially important for experienced users who do not need to listen to the entire message each time the system is used.
9. Provide an introductory or training message to familiarize the user with the system's voice.
10. Do not get caught up in "high-tech fever" — use synthetic speech sparingly and only where it is appropriate and acceptable to the users.

<u>Note</u>. From Chapter 7 in *Human Factors in Engineering and Design* (pp. 191-192) by M. Sanders and E.McCormick, 1987.

structure. This feature enables users to recognize correct options the first time it is heard and should be optional for experienced users. A form of chunking found effective is insertion of pauses at (grammatically) appropriate points. Nooteboom (1983) used pauses in this manner to improve performance with synthetic speech to a level virtually identical with that of natural speech. Waterworth (1983) demonstrated a similar improvement from inserting pauses in a study where subjects recalled automatically generated telephone numbers.

Guidelines provided in Table 1 on page 20 exemplify qualitative guidance provided in current literature. Few, if any, collections of quantitative standards can be found. McKinley, Anderson and Moore (1982) provided an exception by specifying two performance levels used as criteria by the Air Force Aerospace Medical Research Laboratory to evaluate synthetic speech system prototypes. Those criteria require a Modified Rhyme Test score of 80% correct or better and a reaction time of 250 milliseconds (msec) or less. Reaction time used in their criteria measured time from the end of the speech presentation until subject response. This differs from the system response time measure used in this study. However, commercial applications with accuracy ratings of 80% would experience little success.

Database Organization

Despite the large amount of research on optimum menu configurations for visual databases, very little information exists for audible databases. Of the many issues to be resolved in audible databases, perhaps the main issue is the one of organization. Short-term memory and information recall makes menu breadth and depth crucial to the display designer. *Breadth* is number of choices at each menu level and *depth* is the number of menu levels. A 2x6 database like the one used in this study has 2 choices at each level with 6 levels. Snowberry, Parkinson, and Sisson (1983) found subjects performed poorly in searches using 2x6 visual databases and postulated three reasons. First, subjects might have forgotten the target. To counter this factor, Snowberry *et al.* recommend continuous display of the target, a feature of this study's design. Second, subjects may forget the pathway to the target. Since this study assumed infrequent users, the database was designed to make learning a pathway unnecessary. Finally, instead of associating a target with a path of options (the intended searching strategy of Snowberry *et al.*), subjects tended to base selections of options on perceived associations between displayed items and the target. This last explanation posed no problem for this study since an association between menu items and targets was the intended searching strategy for the database.

An additional searching or navigational aid evaluated in this study was use of two voices to speak menus in an alternating fashion. It was thought use of alternating male and female voices would enable a subject to distinguish different levels of the database better and consequently, perform a more efficient (faster) search. Additionally, this voice coding scheme would also assist the subject tracking the depth of menu level progression. Kidd (1982) recommended use of auditory cues such as tones or different voices for just these reasons.

Method

Experimental Design

The experimental design consisted of a 2x2x2 between subjects factorial design. This design as shown in Figure 3 on page 24 contains three independent variables: voice type, coding scheme, and speech rate.

Voice Type and Coding Scheme

Voice type and coding scheme were fixed-effects, between subject variables. Two levels of each variable were fully crossed to create four conditions of voice type and coding scheme. DECtalk's file voice, Perfect Paul, represented the male voice and Beautiful Betty, the female voice. In half of the conditions, either the male or the female voice was used as the sole voice in the synthetic speech display. The remaining conditions employed alternating voices as the subject progressed through

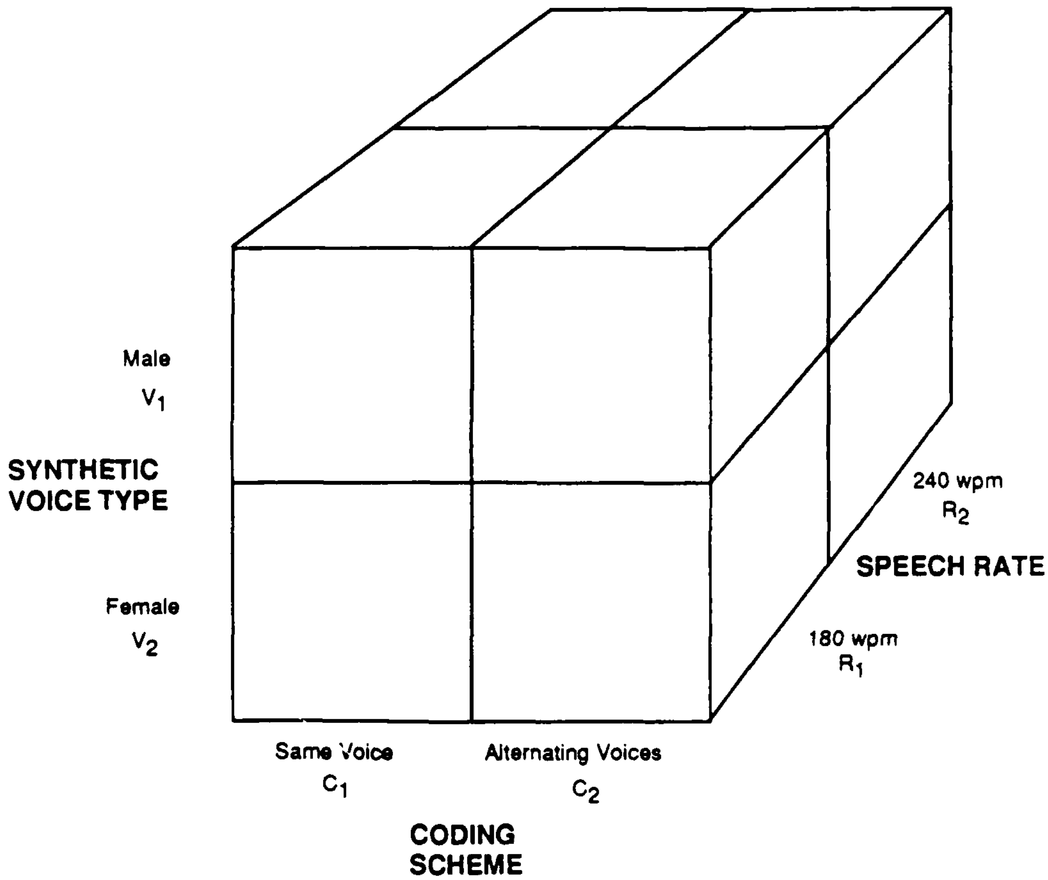

Figure 3. In each condition, 4 subjects searched for 16 targets.

the database levels. In one condition, the female voice began by pronouncing the main menu options followed by the male voice pronouncing the next menu level. This alternating female/male voice pattern continued to the final database level where the information message, a complete sentence, was spoken by the beginning voice — in this case, the female voice. The other condition of alternating voices began and ended with the male voice. This ensured target or information messages in the bottom database level were spoken by both voice types, one in each alternating voice scheme.

Speech Rate

Speech rate was a fixed-effects, between subjects variable. Two levels of this variable were investigated: 180 words-per-minute (wpm), and 240 wpm. Speech rate affected both keywords and information messages which were complete sentences (subject-verb-object). Speech rate was fully crossed with the four conditions of voice type and coding scheme to create the eight treatment combinations depicted in Table 2 on page 26.

Subjects

This study employed 4 subjects in each of 8 treatment combinations of voice type, coding scheme and and speech rate yielding a total of 32 subjects. Volunteers

Table 2. List of Experimental Conditions for 32 Subjects

Condition Number	Treatment Number	Voice Type	Coding Scheme	Speech Rate
1	1	Male	Same	180
	2	Male	Same	180
	3	Male	Same	180
	4	Male	Same	180
2	5	Male	Same	240
	6	Male	Same	240
	7	Male	Same	240
	8	Male	Same	240
3	9	Female	Same	180
	10	Female	Same	180
	11	Female	Same	180
	12	Female	Same	180
4	13	Female	Same	240
	14	Female	Same	240
	15	Female	Same	240
	16	Female	Same	240
5	17	Male/Female	Alternating	180
	18	Male/Female	Alternating	180
	19	Male/Female	Alternating	180
	20	Male/Female	Alternating	180
6	21	Male/Female	Alternating	240
	22	Male/Female	Alternating	240
	23	Male/Female	Alternating	240
	24	Male/Female	Alternating	240
7	25	Female/Male	Alternating	180
	26	Female/Male	Alternating	180
	27	Female/Male	Alternating	180
	28	Female/Male	Alternating	180
8	29	Female/Male	Alternating	240
	30	Female/Male	Alternating	240
	31	Female/Male	Alternating	240
	32	Female/Male	Alternating	240

from the university community were provided monetary compensation for their participation. Average age was 19.9 years with a range from 18 to 27.

Experimental Apparatus

A Beltone 109 Audiometer was used to assess subjects' gross hearing abilities. For the experimental task, Digital Equipment Corporation's (DEC) DECtalk speech synthesizer provided the speech display. Task presentation and data recordings were executed by a VAX 11/750 mainframe system connected to two DEC VT220 terminals using a specially developed PASCAL program. The experimenter station used one VT220 terminal (visual display unit with separate keyboard) to initialize and monitor each session. The subject's station also used a VT220 terminal coupled with a touch-tone speaker phone (Panasonic VA-8205). The telephone's speaker — not the handset — presented the speech display. The volume control was taped over to provide a constant volume level for all subjects. A JVC GX-S700 video camera provided visual and aural monitoring of subjects to video monitors located at the experimenter's station in an adjacent room. Audio or video recordings of experimental sessions were not made.

Information Database

Organization and Keywords

The database constructed for this study contained information about typical department store items. The database was a 2x6 hierarchy containing 6 levels of menus with each menu having 2 items (see Figure 4 on page 29 and Figure 5 on page 30). Each menu item or keyword served as a title for a group of related items (e.g., "entertainment" is a keyword for "music" and "books"). Keywords were selected to allow grouping of store items into sets of 2, 4, 8, 16, and 32, and 64 keywords for each menu level.

Preliminary study efforts attempted to ensure sets of store items were reasonably distinct from each other to reduce searching errors due to semantics or ambiguous keywords. Keywords found by the preliminary study to be grossly unintelligible or consistently misconstrued were discarded and replaced with synonyms or similar items. Manual phoneme or stress polishing was not done to enhance DECtalk pronunciation. However, compound words were entered in an exception dictionary with hyphens at the appropriate location to reduce mispronunciation (i.e., basket-ball, sweat-pants). Finally, contextual clues were provided by the department store scenario to help subjects recognize keywords in both menu levels and information messages.

the database levels. In one condition, the female voice began by pronouncing the main menu options followed by the male voice pronouncing the next menu level. This alternating female/male voice pattern continued to the final database level where the information message, a complete sentence, was spoken by the beginning voice — in this case, the female voice. The other condition of alternating voices began and ended with the male voice. This ensured target or information messages in the bottom database level were spoken by both voice types, one in each alternating voice scheme.

Speech Rate

Speech rate was a fixed-effects, between subjects variable. Two levels of this variable were investigated: 180 words-per-minute (wpm), and 240 wpm. Speech rate affected both keywords and information messages which were complete sentences (subject-verb-object). Speech rate was fully crossed with the four conditions of voice type and coding scheme to create the eight treatment combinations depicted in Table 2 on page 26.

Subjects

This study employed 4 subjects in each of 8 treatment combinations of voice type, coding scheme and and speech rate yielding a total of 32 subjects. Volunteers

Table 2. List of Experimental Conditions for 32 Subjects

Condition Number	*Treatment Number*	*Voice Type*	*Coding Scheme*	*Speech Rate*
1	1	Male	Same	180
	2	Male	Same	180
	3	Male	Same	180
	4	Male	Same	180
2	5	Male	Same	240
	6	Male	Same	240
	7	Male	Same	240
	8	Male	Same	240
3	9	Female	Same	180
	10	Female	Same	180
	11	Female	Same	180
	12	Female	Same	180
4	13	Female	Same	240
	14	Female	Same	240
	15	Female	Same	240
	16	Female	Same	240
5	17	Male/Female	Alternating	180
	18	Male/Female	Alternating	180
	19	Male/Female	Alternating	180
	20	Male/Female	Alternating	180
6	21	Male/Female	Alternating	240
	22	Male/Female	Alternating	240
	23	Male/Female	Alternating	240
	24	Male/Female	Alternating	240
7	25	Female/Male	Alternating	180
	26	Female/Male	Alternating	180
	27	Female/Male	Alternating	180
	28	Female/Male	Alternating	180
8	29	Female/Male	Alternating	240
	30	Female/Male	Alternating	240
	31	Female/Male	Alternating	240
	32	Female/Male	Alternating	240

from the university community were provided monetary compensation for their participation. Average age was 19.9 years with a range from 18 to 27.

Experimental Apparatus

A Beltone 109 Audiometer was used to assess subjects' gross hearing abilities. For the experimental task, Digital Equipment Corporation's (DEC) DECtalk speech synthesizer provided the speech display. Task presentation and data recordings were executed by a VAX 11/750 mainframe system connected to two DEC VT220 terminals using a specially developed PASCAL program. The experimenter station used one VT220 terminal (visual display unit with separate keyboard) to initialize and monitor each session. The subject's station also used a VT220 terminal coupled with a touch-tone speaker phone (Panasonic VA-8205). The telephone's speaker — not the handset — presented the speech display. The volume control was taped over to provide a constant volume level for all subjects. A JVC GX-S700 video camera provided visual and aural monitoring of subjects to video monitors located at the experimenter's station in an adjacent room. Audio or video recordings of experimental sessions were not made.

Information Database

Organization and Keywords

The database constructed for this study contained information about typical department store items. The database was a 2x6 hierarchy containing 6 levels of menus with each menu having 2 items (see Figure 4 on page 29 and Figure 5 on page 30). Each menu item or keyword served as a title for a group of related items (e.g., "entertainment" is a keyword for "music" and "books"). Keywords were selected to allow grouping of store items into sets of 2, 4, 8, 16, and 32, and 64 keywords for each menu level.

Preliminary study efforts attempted to ensure sets of store items were reasonably distinct from each other to reduce searching errors due to semantics or ambiguous keywords. Keywords found by the preliminary study to be grossly unintelligible or consistently misconstrued were discarded and replaced with synonyms or similar items. Manual phoneme or stress polishing was not done to enhance DECtalk pronunciation. However, compound words were entered in an exception dictionary with hyphens at the appropriate location to reduce mispronunciation (i.e., basket-ball, sweat-pants). Finally, contextual clues were provided by the department store scenario to help subjects recognize keywords in both menu levels and information messages.

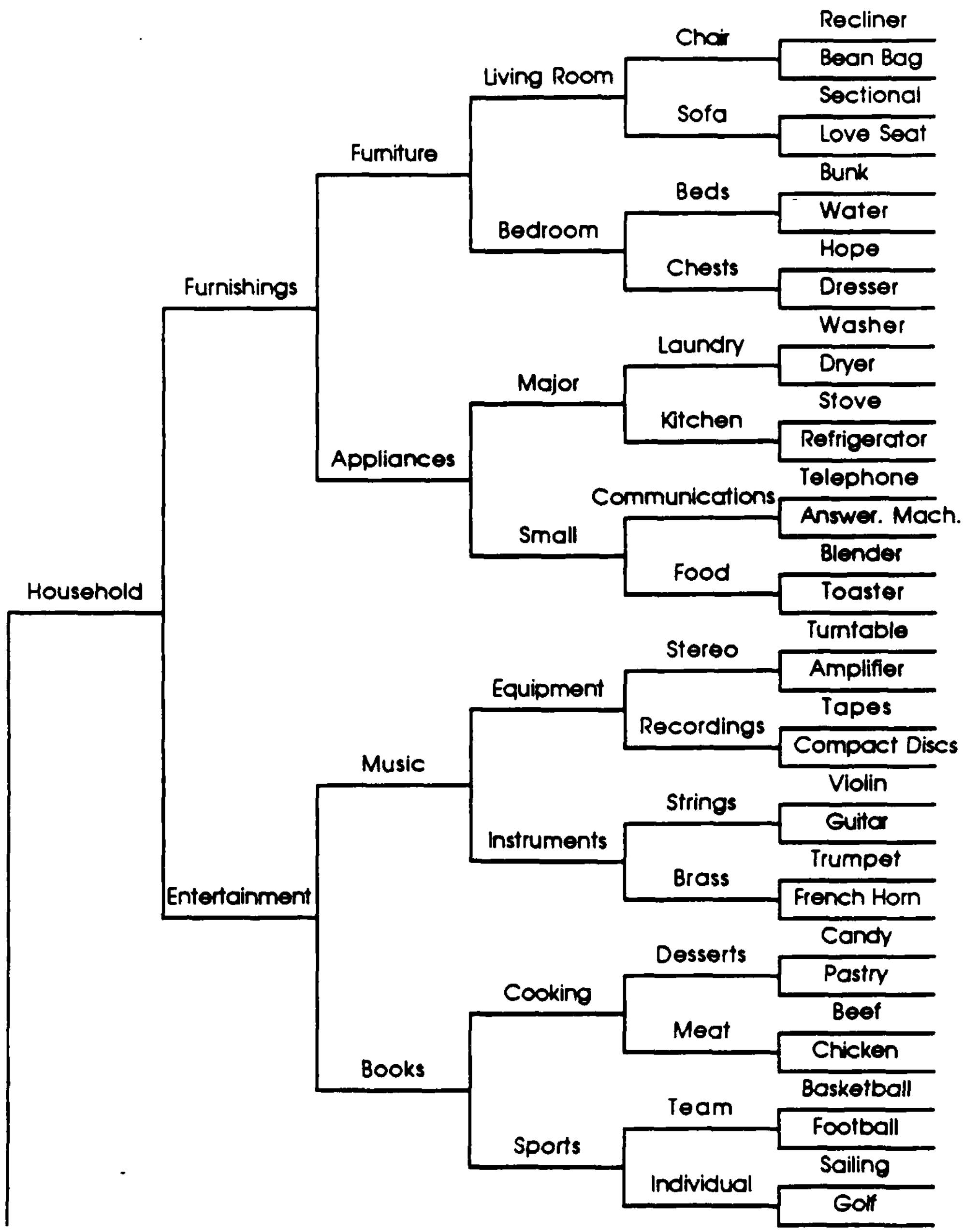

Figure 4. Diagram of the Household-half of the 2x6 hierarchical database.

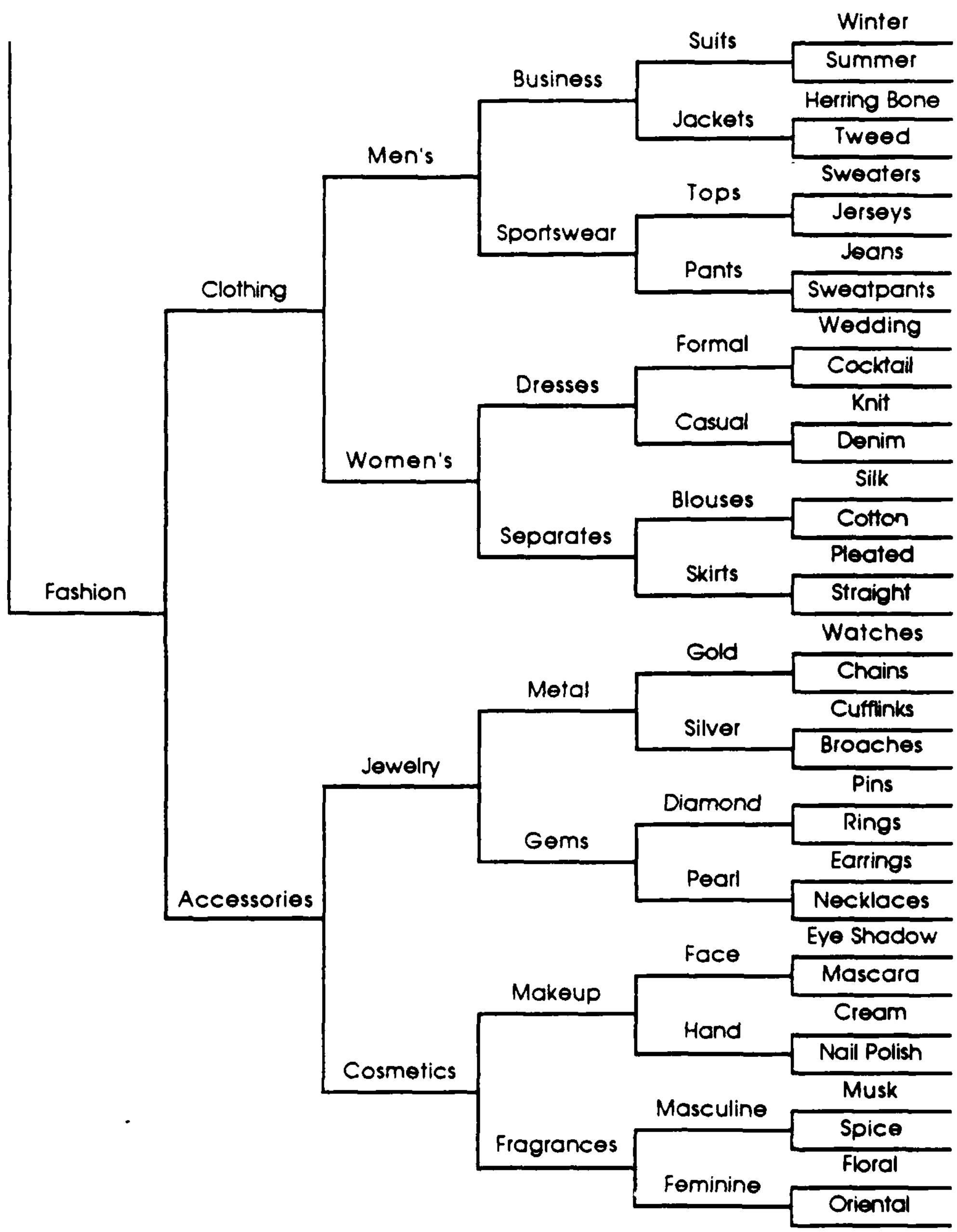

Figure 5. Diagram of the Fashion-half of the 2x6 hierarchical database.

Informati·n Messages

Each of 64 bottom-level keywords functioned as a title for an information message. The messages were of four types: Location, Price, Availability, or Information. Each message had the form of *adjective-noun-verb-preposition-object* (i.e., "Silk blouses are sold for half-price."). As shown in Table 3 on page 32, using a restricted set of verbs and prepositions and non-varying sentence construction standardized the information message format. This standard format made the middle section of each message familiar to the subject and reduced linguistical, context clues as to the meaning of the message. Consequently, the first and last two words in each message could be scored both collectively and separately for transcription accuracy (Merva, 1987). Other guidelines used to construct information sentences are provided in Table 4 on page 33.

Experimental Protocol

Preliminaries

The experimental session began with each subject reading and signing the informed consent form (see Appendix B). Consenting subjects then completed a demographic survey form (see Appendix C). Next, the experimenter administered a hearing test to each subject to eliminate data from "hard of hearing" subjects (American National Standards Institute, 1973). Hearing test criterion was the hearing

Table 3. Information Messages Format

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

Table 4. Rules Used for Developing Information Sentences

1. Each information message was a single sentence.
2. Standard syntax was used for each sentence.
3. Cliches, proverbs, and other stereotyped constructions were avoided.
4. Four message types of information, location, availability, and price were required.
5. Only four words were scored in each sentence.
6. Scored words were never duplicated in any other information message.
7. No proper nouns were allowed as scored words.

of two out of three pulsed tones at 26dB between 750 and 4000 hertz (hz). Subjects unable to pass the test were still allowed to participate, but their data were discarded. This occurred for one subject in this study. After the hearing test, the experimenter used the speakerphone's auto-dial feature to call the department store information system. The synthesizer spoke an introduction and instructions as the subject read along using a written guide (see Appendices D and E). The voice spoke at either 180 wpm or 240 wpm reflecting the subject's assigned treatment condition. The synthesizer used the *dominant* voice for the condition experienced by the subject. For conditions with one voice, the dominant voice was the same voice as heard by the subject in experimental trials. In conditions employing an alternating voice coding scheme, the dominant voice was the voice that spoke the main (or first) menu level and the information message. When the subject completed reading and listening to the instructions, the experimenter played a video tape which repeated the instructions and demonstrated a target search through the database.

Following the instruction tape, the experimenter answered questions and emphasized any differences between the demonstration and conditions the subject was to experience. The experimenter then depressed the space bar on the subject's keyboard causing the synthesizer to review keypad functions available to the subject (see Appendix F). Again, the synthesizer used the dominant voice at the subject's assigned rate.

Experimental Session

The subject then began a practice series of two trials by using the speakerphone to call the department store information system as done earlier. The system "answered" using the dominant synthetic voice to offer a brief review of task instructions. Following this review or a four-second timeout if instructions were not selected by the subject, the first practice *target* was displayed on the computer terminal's display screen for 15 seconds. The first sample target message read, "What is the information about golf books?" At the end of the 15-second display, a "ready..." message displayed on the screen below the target indicated the search was about to begin. The target was displayed on the computer screen throughout the target search. Two seconds after the ready message, a "Begin the search" message was displayed on the screen and the information system spoke the first level menu.

When the subject heard a keyword relating to the target, that keyword was selected by pressing the "#" key on the telephone keypad. The system then responded by speaking the next lower menu level of keywords related to the keyword previously selected. If subjects wanted to backup a menu level, they pressed the "*" key. To return to the main menu, subjects used the "0" key. In this fashion, subjects navigated through the audible database until finding the store item displayed in the target message on the display screen. If the subject arrived at an incorrect store item, the system would speak, "At store item, _______; continue search." To continue the search, subjects depressed the "*" or "0" key.

Upon subject selection of a correct, bottom-level item, the information system requested subjects to depress the "2" key to hear the information message related to the store item. After speaking the information message, the computer screen displayed a message requesting the subject to transcribe the information message just heard. This request replaced the target message displayed during the search. There was no time limit for the transcription task with subjects encouraged to transcribe their best guess if unsure of their answer. After typing in the answer, a series of three computer-displayed messages prompted subjects for subjective ratings (see

Appendix H). The first asked subjects to rate the certainty of their transcription on a scale of 1 (very uncertain) to 7 (very certain). A second bi-polar adjective scale followed the first and asked subjects to rate the difficulty in understanding the message. Again the scale was from 1 (very difficult) to 7 (very easy). Finally, subjects rated difficulty in locating the store item on a scale of 1 (very difficult) to 7 (very easy).

After subjects completed the third rating, a second practice target appeared on the computer screen and as before, fifteen seconds later, the search began by speaking the first menu level. Following this second search, the system hung up and the experimenter offered subjects a rest period. Following the rest period, subjects began the main experimental session by calling the information system as they had done for the two practice searches. Searches proceeded in the same manner as practice trials until the subject found eight targets. After completing the third seven-point scale rating for the eighth target, a "TAKE A BREAK!" message appeared for one minute before another message appeared instructing subjects to press the spacebar to continue. Following the break, subjects completed the remaining eight target searches.

After completion of 16 trials, subjects used the computer terminal to answer 7 additional questions about the telephone information system in the form of seven-point ratings (see Appendix H). Then the experimenter conducted a structured interview of 17 to 21 questions concerning subject impressions of the synthetic voice(s) used in the display and the display application in general (see Appendix I). Subjects receiving an alternating voice condition were asked four questions more (21 total) concerning differences between the two voices used in the display. Finally each subject was debriefed on the experiment's purpose, paid and thanked for their participation. Figure 6 on page 38, illustrates the major portions of each experimental session with average times shown for each portion. Total time for the experimental session ranged from one hour, fifteen minutes to one hour, forty-five minutes, with the average session time per subject lasting approximately one hour, thirty minutes.

Dependent Measures and Data Collection

The experimental task as experienced by a subject was actually two tasks in series: a search task of finding a correct store item followed by a message transcription task of typing the information message into the computer. If a subject arrived at an incorrect store item, the message, "At store item, _______; continue search.", prompted the subject to continue the search until reaching the correct item. Consequently, since searches for a specific store item by all subjects eventually ended at the same store item, this allowed direct comparison of search task measures among subjects. Likewise, since all subjects heard the same, 16 information messages, intelligibility scores could be directly compared as well.

All measures were in the form of keystrokes on the VT220 terminal keyboard or keypresses on the telephone keypad. Both keystroke and keypresses were recorded by a metering package in the software program for the experimental session. Below are 4 objective (performance) and 10 subjective (preference) measures used to assess effects of the independent variables.

Objective Measures

- target search time ratio
- target search efficiency ratio

WELCOME AND ORIENTATION (~ 15 mins)

- Informed Consent
- Subject Information Questionaire
- Hearing Test

INSTRUCTIONS AND PRACTICE (~ 20 mins)

- Introduction (audio - written)
- Instructions (audio - written)
- Video Instructions
- Telephone Key Instructions (audio - written)
- Subject Recapituation of Instructions
- Practice Targets (n=2)

EXPERIMENTAL TASK (~ 30 mins)

- 8 Experimental Targets
  - *Target Search*
  - Transcription
  - Target ratings
- Break (minimum 1 minute)
- 8 Experimental Targets
  - Target Search
  - Transcription
  - Target ratings
- Post Experimental Ratings

POST EXPERIMENTAL SESSION (~ 15 mins)

- Debriefing
- Payment and Dismissal

Figure 6. Outline of Experimental Session Events.

- invalid keypresses
- message transcription errors — strict and synonym

Subjective Measures

After each target search:

- message transcription certainty rating
- message understanding difficulty rating
- search difficulty rating

After completion of all 16 target searches:

- system ease of use
- voice intelligibility
- voice naturalness
- voice speech rate
- system response time
- system input timeout
- menu organization

Search Task Measures

Target search time ratio is an average ratio score of a subject's total search time compared to the minimum search time taken by an expert user. A search time ratio of 1.0 would indicate an "expert" performance by a subject. Expert search time was determined by running a real-time computer simulation of expert searches under conditions experienced by subjects. Each simulation run score was a combination of system time requirements and 0.57 seconds for each menu level selection. This selection time was taken from the American Institutes for Research Data Store (Munger, Smith, and Payne, 1962) for an expert user pressing a pushbutton when cued. Every time a selection was required in a simulation run, this value was used. System time requirement included three values: system response times to user inputs (set at 0 seconds for all 8 treatment conditions), system timeouts or the amount of time provided to users for keypad input (set at 4 seconds for all 8 treatment conditions) and the minimum amount of time the system required to speak the necessary menu items.

However, despite setting the input timeout parameter at 4 seconds, the actual timeout varied by as much as ± 0.5 seconds. This variability was a function of system software. System speech, the third facet of system time requirement, also varied as a function of speech rate and voice type. Because of these small variabilities in DECtalk system time requirements and system response times, average expert scores were obtained. As in the overall experimental design, four real-time simulation runs per condition were conducted to achieve an average expert score for a particular condition. An average search time score for each condition was then combined with the average search time for subjects in the same condition to form the search time ratio score for each subject.

Target search efficiency ratio is a score of subject search efficiency formed by the ratio of minimum number of keywords required to be heard in order to reach a store item to the actual number of keywords heard by a subject. As shown in Table 5 on page 42, target store items were symmetrically distributed among number of keywords required. The total number of keywords each subject heard for all 16 searches was combined with the minimum number of keywords required for all 16 searches. As in target search time ratios, a target search efficiency ratio score of 1.0 would indicate perfect performance by a subject.

Invalid keypresses are keypresses inappropriate at the time of occurrence. Either the key is not defined or a defined key is depressed at an inappropriate time such as depressing the "2" key before reaching an information message. The measure used in this study was the average number of invalid keypresses per search.

Transcription Task Measures

Message transcription errors as calculated in this study is a measure based on a design used by Merva and Williges (1987) to investigate effects of speech rate, message repetition, and information placement on synthesized speech intelligibility. In their scheme, the beginning and end two words of each transcription are checked for accuracy. One point is given for each correct word. Under "strict" scoring, words in the response must be exactly the same as words in the spoken message to be counted as correct. Spelling errors were not counted as incorrect as long as the word remained phonetically correct. "Synonym" scoring allows synonyms for the spoken words to be accepted as correct. Subject responses in this study were scored under both rules. Synonym scoring allows for the variability in human assimilation of spoken words. If a subject transcribes the word, "luggage", for the spoken word, "baggage", one cannot determine if this is due solely to intelligibility

Table 5. Minimum Number of Keywords Required

Keywords Heard	*Number of Target Store Items*
6	1
7	1
8	3
9	6
10	3
11	1
12	1

or includes assimilations effects of comprehension. Synonym scoring effects a compromise for this dilemma by allowing for contextually correct answers.

Hypotheses

The general null hypotheses were different levels of each independent variable or any combination of independent variables would have no effect on the value of any dependent measure. Alternative hypotheses contended an effect but did not suggest a direction. Analysis questions posed by alternative hypotheses are stated in Table 6 on page 44 and Table 7 on page 46.

Table 6. Main Analysis Questions

Cell Comparison	*Task Measures*	*Question*	*Implication of Significance*
Voice (V)	TT	Do scores vary between voices?	Basic intelligibility.
Coding (C)	ST	Do scores improve with measure of information coding?	Efficacy of navigation aids
	TT	Do scores improve with measure of practice?	Possible practice effect if C1 > C2
Speech Rate (R)	ST	Are search scores less with faster rates?	Rate effects on search task performance
	TT	Are less errors made at lower rates?	Rate effects on overall intelligibility
V * C	ST	Do scores vary among combinations of voice type and coding schemes?	Search efficacy of different combinations
	TT	Do scores vary among combinations of voice type and coding schemes?	Effects of practice by same or different voices
V * R	TT	Do scores vary among combinations of voice type and speech rate?	Differential intelligibility as affected by rate
	ST	Do scores vary among combinations of voice type and speech rate?	Search efficacy of different combinations
C * R	ST	Do scores vary among combinations of coding scheme and speech rate?	Search efficacy of different combinations
	TT	Do scores vary among combinations of coding scheme and speech rate?	Differential effects of rate on practice

Note: TT = Transcriptive Task Scores; ST = Search Task Scores

Table 5. Main Analysis Questions (continued)

Cell Comparison	*Task Measures*	*Question*	*Implication of Significance*
V * C * R	ST	Do scores vary among combinations of voice type, coding scheme, and speech rate?	Search efficacy of unique combinations
	TT	Do scores vary among combinations of voice type, coding scheme, and speech rate?	Practice and intelligibility of unique combinations

Table 7. Post Hoc Analysis Questions*

Cell Comparison	*Task Measures*	*Question*	*Implication of Significance*
Are $V_1C_1 > V_1C_2$ and $V_2C_1 > V_2C_2$	TT	Do scores for one condition reflect better performance than another	Practice effect of same voice
If $V_1R_1 = V_2R_1$ or $V_1R_1 > V_2R_1$ and $V_1R_2 > V_2R_2$ or $V_1R_2 < V_2R_2$	TT and ST	Do scores for one condition reflect better performance than another?	Differential intelligibility of voices as affected by rate rate
If $C_1R_1 = C_2R_1$ and $C_1R_2 > C_2R_2$ *or* $C_1R_2 < C_2R_2$	ST	Do scores for one combination of coding scheme and speech rate reflect better performance than another	Differential effect of rate on search efficacy (assuming intelligibility is equal)
Same analysis with corresponding comparisons assuming $C_1R_2 = C_2R_2$			
V * C * R	ST	Are one or more combinations of voice type, coding scheme, and speech rate? better than others?	Search efficacy of unique combinations
	TT	Are one or more combinations of voice type, coding scheme, *and speech rate?* better than others?	Practice and intelligibility of unique *combinations*

* Assumes statistical significance of relevant interactions.

Results

Both performance data from search and transcription tasks and preference data from post-search and post-session ratings were analyzed using descriptive and inferential statistics with data analysis results of $p < 0.05$ considered significant. Dependent measures and data collection procedures are detailed in the Methods Section. Computer files of subject data with manually inserted transcription scores (strict and synonym scored) were input to a data reduction package with reduced data results provided in Appendix J. Statistical data analysis was done with the IBM 370 mainframe computer at Virginia Tech using the Statistical Analysis System (SAS, 1986).

Search Task Data Analysis

A three-way multivariate analysis of variance (MANOVA) for factors of Voice Type, Coding Scheme and Speech Rate was performed for dependent measures of transcription errors (strict and synonym scored), target search time ratios, target search efficiency ratios, and invalid keypresses with results shown in Table 11 on page 52. Conversion of Wilk's *U* criterion to familiar *F* values was used (SAS, 1982) for evaluating overall significance of effects. Means for search task dependent measures categorized by each independent variable are shown in Table 8 on page 49, Table 9 on page 50, and Table 10 on page 51. Speech Rate was the only effect found significant for search task measures, $F\ (5,20) = 3.88$; $p < 0.0128$. The significant overall effect of Speech Rate indicated in Table 11 was not reflected for Speech Rate in subsequent, univariate analyses of variance as shown in Table 12 on page 53, Table 13 on page 54, and Table 14 on page 55.

Scores for target search time ratios ranged from 0.26272 to 0.87358 with a mean of 0.659 or 65.9% of the computer-simulated expert score (see Dependent Measures and Data Collection in Methods Section for dependent measure description). Target search efficiency ratios ranged from 0.33333 to 0.88889 with a mean of 0.74. Invalid keypress averages ranged from 0.0 to 0.3125 with an average score of 0.029. However, only 8 subjects made invalid keypresses with 24 making none. In 3 of the 8 treatment combination cells, no errors were made by any subject (see Appendix J for reduced data listings).

Transcription Task Data Analysis

The three-way multivariate analysis of variance (MANOVA) for factors of Voice Type, Coding Scheme and Speech Rate included analysis of information message

Table 8. Transcription and Search Task Dependent Measure Means by Voice Type

Search Task Measures

Speech Rate	*Search Time Ratio*	*Search Efficiency Ratio*	*Invalid Keypress Average*
Paul	0.67660750	0.75999687	0.01953125
Betty	0.64040062	0.72306000	0.03906250

Transcription Task Measures

Speech Rate	*Strict Errors*	*Synonym Errors*
Paul	9.1250	7.1875
Betty	8.3125	6.1250

Table 9. Transcription and Search Task Dependent Measure Means by Coding Scheme

Search Task Measures

Speech Rate	*Search Time Ratio*	*Search Efficiency Ratio*	*Invalid Keypress Average*
Same	0.63606062	0.71705875	0.03515625
Alternating	0.68094750	0.76599812	0.02343750

Transcription Task Measures

Speech Rate	*Strict Errors*	*Synonym Errors*
Same	8.8125	6.8750
Alternating	8.6250	6.4375

Table 10. Transcription and Search Task Dependent Measure Means by Speech Rate

Search Task Measures

Speech Rate	*Search Time Ratio*	*Search Efficiency Ratio*	*Invalid Keypress Average*
180 WPM	0.68398250	0.75117437	0.02343750
240 WPM	0.63302562	0.73188250	0.03515625

Transcription Task Measures

Speech Rate	*Strict Errors*	*Synonym Errors*
180 WPM	6.6250	4.7500
240 WPM	10.8125	8.5265

Table 11. MANOVA Summary Table for Voice Type x Coding Scheme x Speech Rate Using Search and Transcription Task Measures

Source	*df*	*F**	*p*
Voice Type (V)	5,20	0.60	0.7007
Coding Scheme (C)	5,20	0.39	0.8525
Speech Rate (R)	5,20	3.88	0.0128
V x C	5,20	0.48	0.7867
V x R	5,20	0.70	0.6281
C x R	5,20	0.74	0.6028
V x C x R	5,20	1.22	0.3366

* Approximation of *F* obtained by conversion using Wilk's criterion (SAS, 1986).

Table 12. ANOVA Summary Table for Target Search Time Ratios

Source	df	SS	F	p
Between Subjects				
Voices (V)	1	0.01048750	0.70	0.4100
Coding Scheme (C)	1	0.01611865	1.08	0.3089
Speech Rate (R)	1	0.02077282	1.39	0.2495
V x C	1	0.01381330	0.93	0.3454
V x R	1	0.00501777	0.34	0.5673
C x R	1	0.01016560	0.68	0.4172
V x C x R	1	0.00164494	0.11	0.7427
Subjects/VCR	24	0.35793071		
Total	31	0.43595129		

Table 13. ANOVA Summary Table for Target Search Efficiency Ratios

Source	*df*	*SS*	*F*	*p*
Between Subjects				
Voices (V)	1	0.01091466	0.93	0.3440
Coding Scheme (C)	1	0.01916050	1.64	0.2132
Speech Rate (R)	1	0.00297741	0.25	0.6188
V x C	1	0.00375130	0.32	0.5767
V x R	1	0.00982416	0.84	0.3689
C x R	1	0.00955826	0.82	0.3754
V x C x R	1	0.00220564	0.19	0.6682
Subjects/VCR	24	0.28115277		
Total	31	0.33954470		

Table 14. ANOVA Summary Table for Invalid Keypress Averages

Source	*df*	*SS*	*F*	*p*
Between Subjects				
Voices (V)	1	0.00305176	0.73	0.4019
Coding Scheme (C)	1	0.00109863	0.26	0.6133
Speech Rate (R)	1	0.00109863	0.26	0.6133
V x C	1	0.00012207	0.03	0.8659
V x R	1	0.00982416	0.84	0.3689
C x R	1	0.00598145	1.43	0.2439
V x C x R	1	0.02062988	4.92	0.0362
Subjects/VCR	24	0.10058594		
Total	31	0.33954470		

transcription errors obtained under strict and synonym scoring. Means for transcription task scores are also found in Table 8 on page 49, Table 9 on page 50, and Table 10 on page 51. The significant overall effect for Speech Rate found in the MANOVA also requires further analysis of transcription task dependent measures. Significant effects of speech rate were found in subsequent, univariate analyses of variance as shown in Table 15 on page 57 and Table 16 on page 58 for both strict and synonym scoring. Total transcription errors per subject ranged from 2 to 20 under strict scoring and from 1 to 18 under synonym scoring. Total transcription error means of 8.719 (strict) and 6.656 (synonym) were significantly different with $t\ (31) = 8.69$, $p < 0.0001$.

Transcription Error Analysis by Sentence

Because of observations during data collection and calculation, errors by sentence were analyzed in detail. Total errors made by sentence are depicted in Figure 7 on page 60 in the order each information message sentence was heard by subjects. Additionally, the number of subjects missing each sentence is shown in Figure 8 on page 61. Obviously, sentence 8 and 11 resulted in more errors than others with more subjects making errors for those information message sentences than others. However, the strict and synonym error pattern for sentence 11 differ compared to that of sentence 8. Detailed review of errors for sentence 11 revealed 18 of the 42 strict errors resulted from subjects substituting the word, "samples" for "samplers" which when scored under synonym rules is counted as correct. If errors from these sentences were deleted from the total message transcription errors then total error means would be 5.031 (strict) and 4.062 (synonym). These means, like

Table 15. ANOVA Summary Table for Message Transcription Errors — Strict Scoring

Source	*df*	*SS*	*F*	*p*
Between Subjects				
Voices (V)	1	5.28125	0.35	0.5589
Coding Scheme (C)	1	0.282125	0.02	0.8923
Speech Rate (R)	1	140.28125	9.33	0.0054
V x C	1	0.28125	0.02	0.8923
V x R	1	22.78125	1.52	0.2302
C x R	1	2.53125	0.17	0.6852
V x C x R	1	0.28125	0.02	0.8923
Subjects/VCR	24	360.75		
Total	31	532.46875		

Table 16. ANOVA Summary Table for Message Transcription Errors — Synonym Scoring

Source	*df*	*SS*	*F*	*p*
Between Subjects				
Voices (V)	1	9.03125	0.57	0.4564
Coding Scheme (C)	1	1.53125	0.10	0.7580
Speech Rate (R)	1	116.28125	7.38	0.0120
V x C	1	0.78125	0.05	0.8257
V x R	1	11.28125	0.72	0.4059
C x R	1	0.03125	0.00	0.9649
V x C x R	1	0.03125	0.00	0.9649
Subjects/VCR	24	378.25		
Total	31	517.21875		

those including errors from sentences 8 and 11, are also significantly different with $t(31) = 5.16$, $p < .0001$.

Errors for the first eight sentences were also compared to errors for the last eight sentences to assess training effects. Results were significant for both strict, $t(31) = 4.714$; $p < .0001$, and synonym scoring, $t(31) = 7.602$; $p < .0001$. Means for strict scoring data were 5.531 for the first 8 sentences and 3.188 for the last 8. Means for synonym scoring data were 5.125 for the first 8 sentences and 1.531 for the last 8. Because of these findings, difference scores between the first and last 8 sentences were calculated for each subject (all scores were in the same direction) and analyzed using a three-factor ANOVA procedure. As shown in Table 17 on page 62 and Table 18 on page 63, a significant effect for voice was found for both strict and synonym scoring with subjects showing greater improvement with the male voice (mean = 4.562) than the female (mean = 2.625). As an additional comparison, transcription score means reflected as percent correct are shown by Voice Type in Table 19 on page 64.

Subjective Measures

Median scores were computed in the data reduction program for each subject's transcription certainty, difficulty in understanding the information message, and difficulty in locating the store item (see Appendix I for rating questions). For the seven ratings conducted after the main experimental task was finished, individual ratings were collected. For each of these ten ratings, median or raw scores were analyzed using the Mann-Whitney *U* test. Each test evaluated possible differences

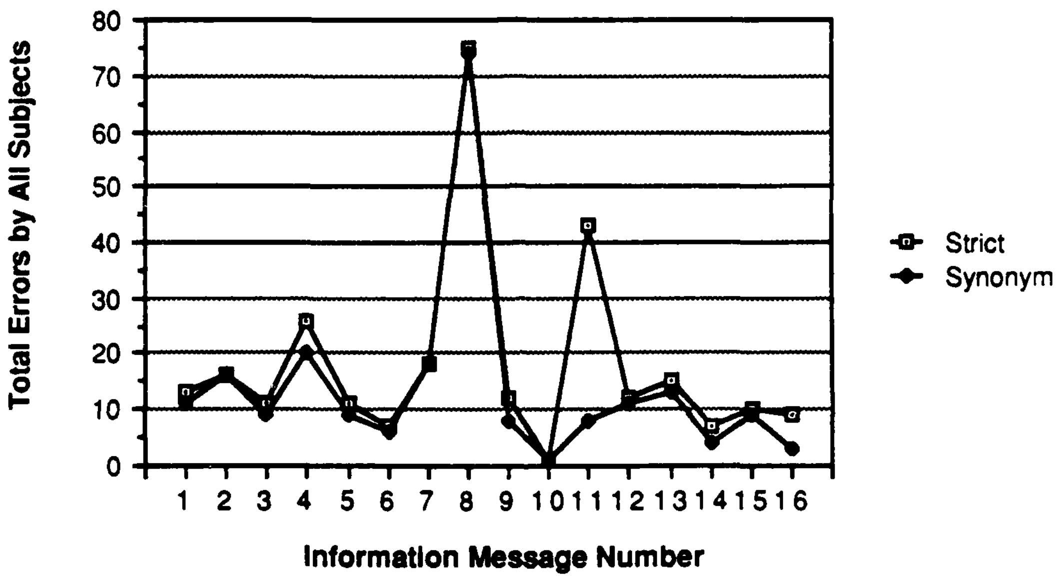

Figure 7. Total errors by information message number.

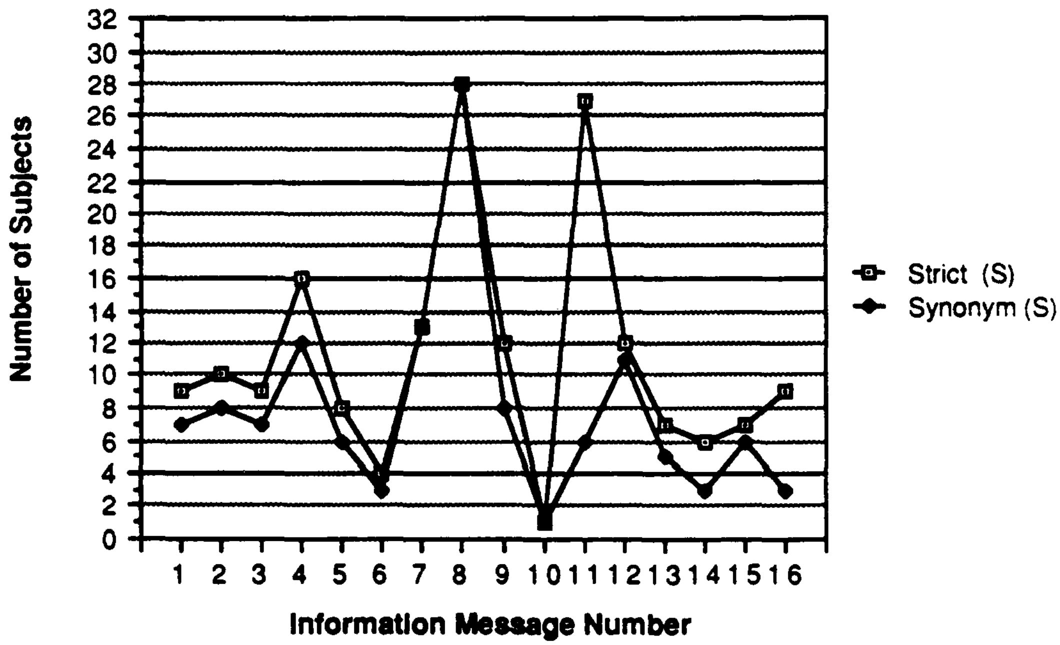

Figure 8. Numbers of subjects missing sentences.

Table 17. ANOVA Summary Table for First 8 - Last 8 Sentence Error Differences — Strict Scoring

Source	*df*	*SS*	*F*	*p*
Between Subjects				
Voices (V)	1	42.781	5.944	0.0226
Coding Scheme (C)	1	1.531	0.213	0.6488
Speech Rate (R)	1	16.531	2.297	0.1427
V x C	1	0.781	0.109	0.7447
V x R	1	5.281	0.734	0.4002
C x R	1	5.281	0.734	0.4002
V x C x R	1	0.281	0.039	0.845
Subjects/VCR	24	172.75		
Total	31	245.217		

Table 18. ANOVA Summary Table for First 8 - Last 8 Sentence Error Differences — Synonym Scoring

Source	*df*	*SS*	*F*	*p*
Between Subjects				
Voices (V)	1	30.031	4.258	0.05
Coding Scheme (C)	1	0.031	0.004	0.9475
Speech Rate (R)	1	16.531	2.344	0.1388
V x C	1	0.281	0.04	0.8434
V x R	1	1.531	0.217	0.6454
C x R	1	3.781	0.536	0.4711
V x C x R	1	0.281	0.04	0.8434
Subjects/VCR	24	169.25		
Total	31	221.717		

Table 19. Mean Percent Correct of Scored Words by Sentence Groups

	All Sentences	First Eight	Last Eight	First Eight*	Last Eight*
Strict Scored					
Paul	85.74	80.27	91.21	88.28	94.92
Betty	87.01	85.16	88.87	91.80	93.56
Synonym Scored					
Paul	88.77	81.64	95.90	89.65	96.48
Betty	90.43	86.33	94.53	92.97	95.51

* Without errors caused by sentences 8 and 11.

between the two levels of each factor of Voice Type, Coding Scheme and Speech Rate. The only significant test occurred for Speech Rate when subjects rated speech rate of the system. Results of all tests are summarized in Table 20 on page 66. A graphical depiction of overall subjective ratings for speech rate as well subject response by independent variable levels is shown in Figure 9 on page 67 and Figure 10 on page 68 respectively. Overall ratings for the remaining nine scales as well as ratings by each independent variable level are depicted in Appendix J.

Table 20. Mann-Whitney U Values* by Factor for Each Subjective Rating Scale

Rating Scale	*Voice Type*	*Coding Scheme*	*Speech Rate*
Median Scored			
Transcription Certainty	114	124	88
Understanding Difficulty	99	115	89
Locating Difficulty	128	112	112
Raw Scored			
Ease of Use	125.5	94	109.5
Voice(s) Intelligibility	118.5	127	84
Voice(s) Naturalness	89	114	123
Speech Rate	106	109.5	28 **
Response Time	123.5	93	107
Input Timeout	86.5	102	124
Menu Organization	97	106	117

* U required for $n_2 = 16$ is 75 for $p < 0.05$ (Siegel, 1956)

** significant for $p < 0.05$

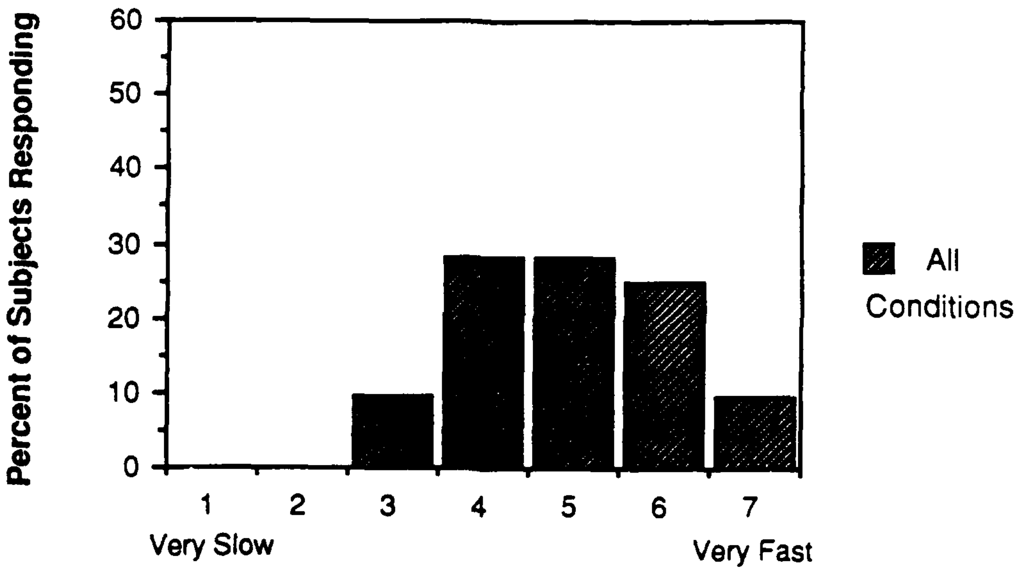

Figure 9. Overall Speech Rate Ratings

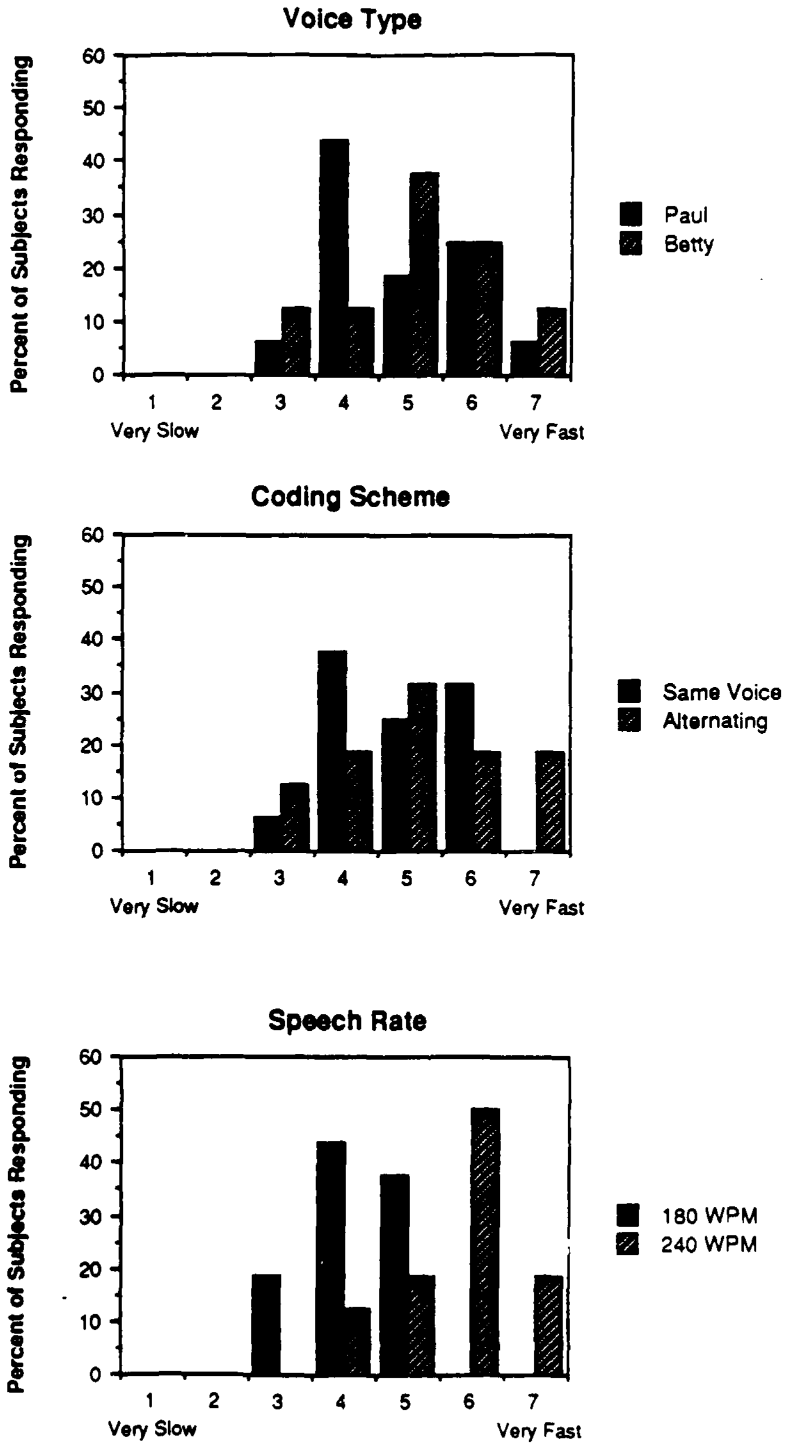

Figure 10. Speech Rate Ratings by Voice Type, Coding Scheme and Speech Rate

Discussion

Performance Results

In this study, alternative hypotheses in the form of questions with associated implications of significance were provided as a framework in which to interpret results. Consequently, Table 6 on page 44 and Table 7 on page 46, which contain these questions for both main and post hoc analyses, will guide the discussion.

Voice

Total transcription scores did not vary between voices for either strict or synonym scoring. However, when transcription scores were analyzed as difference scores between the first eight and last eight sentences, significant effects of Voice Type were found with those hearing Paul showing more improvement in the last eight sentences. Each of these findings are discussed in turn.

Researchers have often reported performance measures such as percent correct for a voice type without reporting statistical significance of their findings. The study by Green, *et al.* (1984) is just such an example. An exception is Pratt's (1987) study comparing four DECtalk voices (including Paul and Betty) to other synthesizers. Percentages were provided as in other studies but statistical analyses (ANOVA and Newman-Keuls) were performed on preference measures. Since statistical significance of Voice Type performance differences is rarely reported, direct comparison of this study's lack of significant difference is not possible.

However, comparison of percentage scores is possible. Transcription accuracy means reported in percent correct (see Table 19 on page 64), differ slightly in relative magnitude from those reported in the literature. Using a sentence transcription task (Harvard Psychoacoustic Sentences) analogous to one used in this study, Green, *et al.* reported percentages of 95.3% for Paul's voice and 90.5% for Betty's. Results found in this study (strict scoring for comparison with Green, *et al.* (1984) study) show 85.74% for Paul and 87.01% for Betty, which are similar performance levels when compared to Green, *et al.*. However, this comparison and others must consider at least three differences between the two studies: first, Green, *et al.* used an earlier version of the DECtalk speech synthesizer (DECtalk version 1.8 for the Green, *et al.* study compared to the DECtalk version 2.0 used in this one); second, the task required of subjects differed substantially between the two studies — simple transcription of synthetically spoken sentences (Green, *et al.*) as compared to the integrated task (search and transcription) required by simulation of a telephone information system; finally, lower percentages reported in this study probably reflect scoring of the four most difficult words in the sentence as compared to Green, *et al.*'s procedures of scoring all words in a sentence.

When errors are analyzed as percentage correct for first eight sentences heard and last eight sentences heard, interesting performance results between voices are shown (again, see Table 19 on page 64). Researchers have usually reported a slight performance advantage for DECtalk's Paul voice when compared to the Betty voice (although presence or lack of statistically significant differences are rarely reported thus limiting the power and extent of possible comparisons). However, the percentage correct in the first eight sentences for those hearing Betty's voice is greater than for those hearing Paul's voice. This numerical advantage for Betty disappears in the last eight sentences heard with those hearing Paul averaging 91.21% correct and those hearing Betty, averaging 88.87%.

A finding not previously reported in literature occurred when analysis of transcription scores divided into scores for first eight and last eight yielded a significant difference for both strict and synonym scoring. When difference scores between the first and last eight sentences heard were analyzed, a significant difference for Voice Type was found. Those subjects hearing Paul trained at a significantly faster rate although they began at an apparently (no significant difference) lower level of performance than those hearing Betty. Though this finding demonstrated the effect of training in synthetic speech reported by several researchers including Schwab, Nusbaum and Pisoni (1985), Rosson (1985), and Merva and Williges, (1986), none have mentioned differences observed by Voice Type.

Coding Scheme

Search task scores did not improve (or deteriorate) by using an alternating voice coding scheme nor did transcription task scores reveal a differential practice effect. It was hypothesized those hearing the same voice would have more practice with that voice and consequently perform better on the transcription task. Continuing this reasoning, those experiencing the alternating voice coding scheme would have less practice with the voice used for the transcription task — approximately 50% less — and therefore perform poorly when compared to those experiencing the same voice coding scheme. Therefore, the training effect observed for synthetic speech displays appears to be nonspecific since performance improvement occurs even when different synthetic voices are used in the training session. Alternating voice coding schemes were also intended as a navigation aid enabling subjects to track menu levels more accurately. Results do not support either position, though.

In fact, little research exists on aids for auditory database navigation. Calls for using navigation aids such as the one employed in this study are based more so on intuition than empirical validation (Kidd, 1982). One subject provided an insight to this issue during the debriefing by maintaining he had heard only one voice even though he was assigned to an alternating voice condition. Though most assigned to alternating voice condition acknowledged hearing two voices, many did not think this was an aid to database navigation with some unsure of the pattern of voice alterations. Perhaps instructing subjects on the alternating voice coding scheme would have enhanced its effect. Other reasons for lack of significant findings, for this variable are considered in the discussion of interaction effects and post hoc analyses.

Speech Rate

Speech Rate significantly affected both search task and transcription overall task performance which is consistent with findings from previous studies. However, a more focused effect for Speech Rate was not detected in the three subsequent ANOVA procedures using search task dependent measures. It is possible for a MANOVA procedure to reveal a significant effect when separate ANOVAs do not. This phenomenon reflects superior experimental power of the MANOVA procedure over use of separate ANOVAs when a significant effect is spread across more than one dependent measure (Finkelman, Wolf, and Friend, 1977). As discussed in the Literature Review Section, earlier research has found effects of Speech Rate to be at least consistent if not uniformly significant. And overall results of this study remain consistent with findings of earlier research. Yet, at a speech rate of 240 wpm, intelligibility of synthetic speech does not seem to affect search and transcription tasks equally. Transcription task measures were significant for both MANOVA and ANOVA procedures whereas search task measures were not.

A possible explanation of the lack of focused speech rate effects on search task measures comes from information theory. As posed by Luce, *et al.*, (1983), synthetic speech is thought to increase the cognitive load on the listener as compared to comprehension of natural speech. Regardless of the information theory model considered (serial, parallel or hybrid), this increased cognitive load diminishes capacity in working or short-term memory. Increasing speech rate should further increase the high cognitive load (as compared to natural speech) imposed by synthetic speech, yet no differential effect of Speech Rate was observed for search task measures. Though keywords had a slightly, shorter pronunciation duration at 240 wpm, the 4-second timeout probably enabled subject performance comparable to that observed at 180 wpm. With a 4-second timeout (provided for both 180 and 240 wpm conditions), subjects had time to rehearse and comprehend a keyword prior to the next keyword being presented. This rehearsal time was enough to overcome the diminished cues provided by an assumed poor quality speech signal presented at the high rate of speed.

Under both strict and synonym scoring, Speech Rate significantly affected transcription accuracy, a finding well established in the literature. However, the contribution of rate to this finding may not be just a function of rate. A majority of subjects during debrief described an interfering effect of hearing the phrase, "Begin Transcription", after the information message. Some subjects could be heard repeating the message repeatedly until typing it into the computer. One subject in a 240 wpm condition actually began typing before the computer terminal display had changed as a strategy to preclude forgetting the message because of the Begin Transcription phrase. To borrow again from information theory, this phrase interfered with the critical role of rehearsal required to maintain information in short term memory. At higher speech rates, subjects have less time for rehearsal thus increasing capacity demand of short term memory. The Begin Transcription phrase probably caused an over demand or overload for some subjects' short term memory.

As mentioned in the Results Section, 2 sentences accounted for considerably more errors than the other 14 although error patterns as depicted in Figure 7 were different between sentences 8 and 11. Sentence 8 contained words obviously unintelligible, but subjects hearing sentence 11 could comprehend the meaning if not record the precise words spoken. The most common error for sentence 11 was substitution of the word, "samples" for "samplers". A limited analysis of transcription errors which discarded errors caused by sentences 8 and 11 revealed little differences between earlier analyses containing those errors. However, implications for a designer of synthetic speech displays are clear and point to the need for careful screening of messages with a large number of potential users.

Interaction Effects and Post Hoc Analyses

The same question was posed for all cell comparisons: do scores vary among combinations of independent variable levels? MANOVA results for both search and transcription task dependent measures provided a negative reply to this question. Two possible reasons exist for this negative reply: first, failure to reject the null hypotheses suggest these independent variables hold no import (statistical or pragmatic) for synthetic speech displays; or perhaps these issues could (or do) make a difference but conduct of the experimental study precluded that discovery. For the second reason, several detailed explanations exist.

Dependent measures used in this study could possibly have been insensitive to additional differences caused by manipulation of independent variables. This insensitivity could result from use of dependent measures inappropriate to the dependent variable construct being measured. Effects of independent variables on dependent measures such as search time, search efficiency, and invalid keypresses have not been widely explored. Though an overall effect of speech rate was detected for search task dependent measures, no discrete effects (as reflected by individual ANOVA procedures) were revealed. And it is possible search task measures used in this study were not sensitive enough to detect effects of voice type or coding scheme. The dependent measure of invalid keypresses exemplifies this viewpoint. Out of 32 subjects, 24 never made an invalid keypress with 4 subjects making one invalid keypress, 3 subjects making 2, and 1 subject making 5. In the 8 treatment conditions, 3 had no subjects making an invalid keypress with 2 more conditions having one subject each.

Another reason for possible insensitivity of dependent measures is the strong context provided by the department store setting. Strong contextual clues could have masked possible aiding or debilitating effects of the independent variables. Though keyword intelligibility may have been diminished, the hierarchical relationship of keywords to each other within the limits of a department store settings may have provided the clues needed to overcome a supposedly poorer speech signal. Evidence for this view comes from debriefing comments when a subject explained his search strategy as being a "rule-out" approach. He understood one keyword, "Household" but not the other, "Fashion". Consequently, he chose the keyword, Fashion, whenever the target store item appeared not to fit under the category of Household ("ruling out" the understood keyword). Such a strategy indicated use of broad, contextual clues.

Finally, training provided subjects may have made them less sensitive to variables manipulated in the study and hence, the dependent measures used to assess independent variable effects. Subjects were provided with various forms of training to include two practice runs. This procedure resulted from preliminary studies out of concern that errors generated from the first several searches might reflect task uncertainty as opposed to effects of independent variables. Providing thorough instructions and practice was intended to stabilize measures, not mute them. Again, debriefing comments provide some support as several subjects said they understood the task after the tape though practice runs following the tape were helpful.

If independent variable effects were indeed obscured by insensitive dependent measures, several corrections could be made based on reasoning offered here. First, the number of subjects could be increased resulting in a more powerful test by reducing effects of subject variability. Secondly, subject training could be diminished to more closely resemble naive users and thus possibly render the dependent measures more sensitive. However, careful design of experimental procedures would be necessary to preclude measuring task uncertainty as opposed to the true effects of the paradigm's independent variables. Finally, by decreasing the amount of training, context familiarity is also lowered making intelligibility of keywords (and the effects of independent variables on them) more critical.

Preference Results

Statistical analysis of subjective ratings provided only one significant finding: those subjects assigned to different Speech Rate conditions rated Speech Rate differently and reflected the condition assigned to them. Earlier research consistently supports this finding making Speech Rate a pervasive and strong factor in synthetic speech perception. No further, statistically significant differences between subject groups (classified by independent variable levels) were found. However, in absence of performance data or statistically significant data of any kind, preference or subjective data serve designers as starting points for field trials. Subjective data gathered in this study could perform the same function for a telephone information system with major impressions summarized below.

The majority of subjective ratings provided by subjects were "favorable" to the system. Most subjects tended to be very certain about their transcription accuracy though ratings were not as high for understanding the information message. High ratings given to locating store item difficulty reflect study results of no significant differences found using search task measures. Also, most thought the information system easy to use, possibly a reflection of experimenter-provided training discussed earlier in this section. Ratings for intelligibility and naturalness show some of the more symmetrical distribution of ratings observed with the overall rating for naturalness resembling a normal (Gaussian) distribution centered on a rating of four. Of all ratings, intelligibility and naturalness seemed to be rated lower than other dimensions. Most thought system response time was very fast with ample time (input timeout) to respond. The majority rated menu organization as very simple, a rating which corresponds with subject ratings of very easy in difficulty of locating store items.

Conclusions

The study results imply the following guidelines for use of synthetic speech displays in telephone information systems:

- Use of a 180 wpm speech rate yields better transcription accuracy (intelligibility) as compared to using a speech rate of 240 wpm.
- Use of different speech rates significantly affects search tasks in auditory databases though precise effects are not yet known. Consequently, though designers of synthetic speech displays may desire acceleration of search tasks, use of speech rates faster than 180 wpm needs further research.
- Users are both aware of and sensitive to speech rate.
- When applications require strict or precise recall of spoken utterances, the messages should be screened by a sample of the intended user population to ensure substitutions are absent or at acceptable levels.
- Although using one voice type (male as opposed to female and as represented by DECtalk's Perfect Paul and Beautiful Betty) over another provides no statistically significant advantage, designers should consider training time available to

users as those using the male synthetic voice improve at a faster rate when compared to the female voice.

- Use of alternating voices as a navigation aid in auditory databases provides no apparent benefit.
- Avoid placing phrases not part of an information node immediately following an information message node. Violation of this principle could cause interference in a user's cognitive rehearsal — a process necessary for short term memory retention.

Future research in using synthetic speech displays in telephone information systems hold many questions among which are the following:

- How do training rates between male and female voices (as represented by Paul and Betty) compare? Do listeners of Paul continue to improve at a faster rate while those hearing Betty asymptote in their performance? Do findings support adaptive rate features (user selected or system provided)?
- Does the midband filter function inherent in telephone communication affect synthetic speech performance in ways different from speech heard without using a telephone? Does synthetic speech performance in a telephone display using previous synthetic speech measures (open and closed MRT, Haskins and Harvard sentences) differ from previous results?
- How may search task dependent measures be rendered more sensitive to effects of speech rate and other variables? Do larger number of subjects render the same dependent measures more sensitive? Would field studies reveal differences opposite to findings of laboratory studies? Would search task dependent

measures different from those used in this study reflect performance differences for independent variables used in this study?

- How does synthetic speech rate specifically affect search tasks in auditory databases? Can speech rate (or the achieved effect by decreasing keyword pronunciation duration and timeout rate) be increased for menus as compared to information message nodes?
- Do database organizations other than the formal, hierarchical structure featured in this study offer better performance? For example, does using a database containing more than one path to an information node result in more efficient searches?
- What is the minimum time necessary between an information message node and subsequent system speech in order to prevent interfering with short term memory retention of the information message?
- Are users different where synthetic speech is concerned? Does performance and preference of telephone information systems employing synthetic speech systematically vary along dimensions of the users? What are those dimensions?

Despite its coarticulation problems and lack of sophisticated prosody, synthetic speech at current technological levels remains a viable, auditory display for telephone information systems. Much research is needed though, on auditory database construction and use of synthetic speech in such databases. Research recommendations provided above are in no way exhaustive of auditory display problems pertinent to telephone information studies.

References

Allen, J. (1981). Linguistic-based algorithms offer practical text-to-speech systems. *Speech Technology, 1*, 12-16.

American National Standards Institute (1973). *Psychoacoustic Terminology*. New York: Author.

Atkinson, R.C. and Shiffrin, R.M. (1968). Human memory: a proposed system and its control processes. In K.W. Spence and J.T. Spence (Eds.), *Advances in the psychology of learning and motivation research and theoryd*, 2. New York: Academic Press.

Baddeley, A.D., Thomson, H., and Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of verbal learning and learning behavior, 14*, (575-589).

Barnwell, T.P. (1982, March). On the standardization of objective measures for speech quality testing. In D. Pallett (Ed.), *Proceedings of the Workshop on Standardization for Speech I/O Technology* (pp. 193-200). Gaithersburg, MD: National Bureau of Standards.

Bristow, G. (1984). The speech revolution. In G. Bristow (Ed.) Electronic Speech Synthesis (pp. 7-13). New York: McGraw-Hill.

Broadbent, D.E. (1958). *Perception and communication*. London: Pergamon.

Broadbent, D.E. (1971). *Decision and stress*. New York: Academic Press.

Broadbent, D.E. (1982). Task combination and selective intake of information. *Acta Psychologica, 50*, 253-290.

Conrad, R. and Hull, A.J. (1964). Information, acoustic confusion and memory span. *British Journal of Psychology,55*, 429-432.

Cooper, M. (1987). Human factor aspects of voice input/output. *Speech Technology* 82-86.

Deatherage, B.H. (1972). Auditory and other sensory forms of information presentation. In H.P. Van Cott and R.G. Kinkade (Eds.), *Human engineering guide to equipment design* (Rev. ed.). Washington, D.C.: U.S. Government Printing Office.

Egan, J.P. (1948). Articulation testing methods. *Laryngoscope, 58*, 955-991.

Fairbanks, G. (1958). "Test of phonemic differentiation: the rhyme test." *Journal of Acoustical Society of America, 30*, (7), 596-600.

Finkelman, J.M., Wolf, E.H., and Friend, M.A. (1977). Modified discriminant analysis as a multivariate post-comparison extension of MANOVA for interpretation of simultaneous multimodality measures. *Human Factors, 19*, (3), 253-261.

Flanagan, J.L. (1972). *Speech analysis, synthesis, and perception* (2nd ed). New York: Springer Verlag.

Greene, B.G., Manous, L.M., and Pisoni, D.B. (1984). Perceptual evaluation of the DECtalk: a final report of version 1.8. *Research on Speech Perception, Progress Report Number 10*, Indiana University.

House, A.S., Williams, C.E., Hecker, M.H.L., and Kryter, K.D. (1965). Articulation-testing methods: consonantal differentiation with a closed-response set. *Journal of Acoustical Society of America, 37 (1)*, 158-166.

Kaplan, G. and Lerner, E.J. (1985). Realism in synthetic speech. *IEEE Spectrum, 22*, (4), 32-37.

Kantowitz, B.H. (1974). Double stimulation. In B.H. Kantowitz (Ed.), *Human information processing*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kantowitz, B.H., and Sorkin, R.D. (1983). *Human factors: understanding people-system relationships*. New York: McGraw-Hill.

Kinsbourne, M. (1981). Single-channel theory. In d. Holding (Ed.), *Human skills*. New York: Wiley.

Kidd, A.L. (1982). Problems in man-machine dialogue design. In *Institute of Electrical and Electronic Engineers (IEEE) Proceedings of Telecommunications Conference* (pp. 531-536). Zurich.

Klatt, D.H. (1986). Text to speech: present and future. In *Proceedings of Speech Tech '86*, 221-226. New York: Media Dimensions, Inc.

Labrador, C. and Pai, D. (1984). Experiments in speech interactions with conventional data services. In *Proceedings of Interact '84: 1st IFIP Conference on Human-Computer Interaction* (pp. 104-108). London.

Lane, D. (1981). Attention. In W.C. Howell and E.A. Fleishman (Eds.), *Human performance and productivity, 2.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Luce, P.A., Feustel, T.C., and Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors, 24,* 17-32.

McCormick, E.J. and Sanders, M.S. (1982). *Human Factors in Engineering and Design* (5th Ed.). New York: McGraw-Hill.

McHugh, H.M. (1986). Telephone accessed speech systems and the information explosion. In *Proceedings of Speech Tech '86,* 40. New York: Media Dimensions, Inc.

McKinley, R.L., Anderson, T.R., and Moore, T.J. (1982). Evaluation of speech synthesis for use in military noise environments. In D. Pallet (Ed.), *Proceedings of the Workshop on Standardization for Speech I/O Technology* (pp. 241-245). Gaithersburg, MD: National Bureau of Standards.

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

Merva, M.A., and Williges, B.H. (1986). The effect of contextual knowledge and speech rate on message intelligibility. Unpublished manuscript, Human-Computer Interaction Laboratory, IEOR Department, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Merva, M.A., and Williges, B.H. (1987). The intelligibility of synthesized speech in data inquiry systems. Unpublished manuscript, Human-Computer Interaction Laboratory, IEOR Department, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Miller, G.A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63,* 81-97.

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

Nooteboom, S.F. (1983). The temporal organisation of speech and the process of spoken-word recognition. *IPO (Eindhoven) Progress Report, 18,* 32-36.

Nusbaum, H.C. Dedina, M.J., and Pisoni, D.B. (1984). Perceptual confusion of consonants in natural and synthetic CV syllables. *Research on Speech Perception, Progress Report No. 10,* Indiana University, Bloomington, IN, pp. 153-168.

Nye, P.W., and Gaitenby, J. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Laboratories Status Report on Speech Research, *38,* 169-190.

O'Malley, M.H., and Caisse, M. (1987). How to evaluate text-to-speech systems. *Speech Technology, 3,* (4), 66-75.

Osgood, C.E., Suci, G.L., and Tannenbaum, P.H. (1957). *The Measurement of Meaning.* University of Illinois Press.

Pisoni, D.B., and Hunnicut, S. (1980). Perceptual evaluation of MITalk: the MIT unrestricted text-to-speech system. In proceedings *1980 IEEE International Conference on Acoustics, Speech and Signal Processing Society,* (pp. 572-575). Denver, CO.

Pisoni, D.B. (1982). Speech Technology: the evolution of computers that speak...and listen. *Distinguished Faculty Research Lecture,* February 22, 1982, Indiana University, Bloomington, IN: Office of Research and Graduate Development.

Pratt, R.L. (1987). Quantifying the performance of text-to-speech synthesizers. *Speech Technology, 3,* (4), 54-64.

Rosson, M.B. (1985). Listener training for speech-output applications (Research Report RC 11029, #49529). Yorktown Heights, NY: IBM Watson Research Center.

Rosson, M.B., and Cecala, A.J. (1985). An analysis of listener's reactions to synthetic voices (Research Report RC11398 #51318). Yorktown Heights, NY: IBM Watson Research Center.

Sanders, M.S., and McCormick, E.J. (1987). *Human factors in engineering and design* (6th ed.). New York: McGraw-Hill.

SAS Institute Inc. (1986). *Statistical Analysis System, Release 5.16.* Cary, NC: SAS Institute Inc..

Schmidt-Nielsen, A. (1985). Problems in evaluating the real-world usability of difital voice communication systems. *Behavior Research Methods, Instruments, and Computers, 17* (2), 226-234.

Schwab, E.C., Nusbaum, H.C., and Pisoni, D.B. (1985). Some effects on training on the perception of synthetic speech. *Human Factors, 27,* 3985-408.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill.

Simpson, C.A. (1983). Evaluating computer speech devices for application. In J.C. Warren (ed.), *Proceedings of the Seventh West Coast Computer Faire* (pp. 395-401).

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

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

Simpson, C.A., and Williams, D.H. (1975). Human factors research problems in electronic voice warning system design. In *Proceedings of the 11th Annual Conference on Manual Control* (NASA TMX-62, 464) (pp. 94-106). Moffett Field, CA: NASA-Ames Research Center.

Simpson, C.A., and Williams, D.H. (1980). Response time effects of alerting tone and semantic context for synthesized voice cockpit warnings. *Human Factors, 22,* 319-320.

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

Snowberry, K., Parkinson, S., and Sisson, N. (1983). Computer display menus. *Ergonomics, 26,* 699-712.

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

Voiers, W.D. (1977). Diagnostic acceptability measure for speech communications systems. In *Conference Record, ICASSP '77,* New York: Institute of Electrical and Electronics Engineers.

Voiers, W.D. (1983). Evaluating processed speech using the Diagnostic Rhyme Test. *Speech Technology, 1,* (4), 30-39.

Waterworth, J.A. (1983). Effect of intonation form and pause duration of automatic telephone number announcements on subjective and memory performance. *Applied Ergonomics, 14,* 39-42.

Waterworth, J.A., and Lo, A. (1984). Examples of an experiment: evaluating some synthesizers for public announcements. In A. Mond (Ed.), *Fundamentals of Human-Computer Interaction* (pp. 89-102). London: Academic Press.

Wheale, J. (1980). Pilot opinion of flight deck voice warning systems. In D. Oborne and J. Levis (Eds.), *Human factors in transport research, 1.* New York: Academic.

Appendix A. References Used in Figure 1

Carlson, R. and Granstrom, B. (1975). A phonetically oriented programming language for rule description of speech. In G. Fant (Ed.) *Speech Communication, 2* (pp. 245-253). Stockholm: Almqvist and Wiksell.

Carlson, R. and Granstrom, B. (1976). A text-to-speech system based entirely on Rules. In *Proceedings of the Institute of Electrical and Electronic Engineers (IEEE) Conference on Acoustic and Speech Signal Processes (ICASSP), 1976*, (pp. 686-688).

Coker, C.H., (1967). Speech synthesis with a parametric articulatory model. In J. Flanagan and L. Rabiner (Eds.), *Speech Synthesis*, (pp. 135-139). Stroudsburg, PA: Dowden, Hutchinson, and Ross.

Coker,C.H., Umeda, N., and Browman, C.P. (1973). Automatic synthesis from ordinary english text. *IEEE Audio and Electroacoustics, AU-21*, 293-397.

Cooper, F.S., Liberman, A.M., and Borst, J.M. (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. In *Proceedings of the National Academy of Science, 37*, (pp. 318-325).

Dixon, R.N., and Maxey, H.D. (1968). Terminal analog synthesis of continuous speech using the diphone method of segment assembly. *IEEE Trans. Audio and Electroacoutics, AU-16*, 40-50.

Dudley, H., Riesz, R.R., and Watkins, S.A. (1939). A synthetic speaker. *J. Franklin Institute, 227*, 739-764.

Fant, G. (1953). Speech communication research. *Ing. Vetenskaps Akad., 24*, 331-337. Stockholm.

Fujimura, O. and Lovins, J. (1978). Syllables as concatenative phonetic elements. In Bell and Hooper (Eds.), *Syllables and Segments*. New York: North-Holland.

Gagnon, R.T. (1978). Votrax real time hardware for phoneme synthesis of speech. In *Proceedings of the ICASSP-78.*

Hertz, S. (1982). From text to speech with SRS. *Journal of the Acoustic Society of America, 72*, 1155-1170.

Holmes, J., Mattingly, I., and Shearme, J. (1964). Speech synthesis by rule. *Language and Speech, 7*, 127-143.

Kelly, J., and Gerstman, L. (1961). An artificial talker driven from phonetic input. *Journal of the Acoustic Society of America, 33*, 33.

Klatt, D.H. (1970). Synthesis of stop consonants in initial position. *Journal of the Acoustic Society of America, 47*, 93.

Lawrence, W. (1953). The synthesis of speech from signals which have a low information rate. In W. Jackson (Ed.), *Communication Theory*, (pp. 460-469). London: Butterworths Science Publishers.

Liberman, A., Ingemann, F., Lisker, L., Delattre, P., and Cooper, F. (1959). Minimal rules for synthesizing speech. *Journal of the Acoustic Society of America, 31*, 1490-1499.

Mattingly, I. (1968). Synthesis-by-rule of general American English. *Supplement to Status Report on Speech Research* New Haven, CT: Haskins Laboratories.

Olive, J.P. (1977). Rule synthesis of speech from diadic units. In *Proceedings of the ICASSP-77* (pp.568-570).

Peterson, G., Wang, W., and Sivertsen, E. (1958). Segmentation techniques in speech synthesis. *Journal of the Acoustic Society of America, 30*, 739-742.

Potter, R.K., Kopp, G.A., and Green, H.C. (1947). *Visible speech*. New York: van Nostrand Company.

Rosen, G. (1958). A dynamic analog speech synthesizer. *Journal of the Acoustic Society of America, 30*, 201-209.

Stevens, K.N., Kasowski, S. and Fant, G. (1953). An electrical analog of the vocal tracts. *Journal of the Acoustic Society of America, 25*, 734-742.

Appendix B. Participant's Informed Consent Form

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

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

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

3. You have the right to be informed of the overall results of the experiment. If you wish to received 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 David W. Herlong. He 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) 961-4602

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

Charles Waring
Office of Sponsored Research Programs
301 Burruss Hall
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061
(703) 961-5283

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

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

____

Participant's Signature

____

Witness' Signature

____

____

____

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

Appendix C. Subject Information Questionnaire

Age: _____ Sex: _____ Native language: ___________________

Please list any hearing impairments you may have:

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

1. How experienced are you with using computers?

No experience | Some experience | Experienced | Very Experienced

2. How experienced are you with using information systems?

No experience | Some experience | Experienced | Very Experienced

3. How experienced are you with listening to synthesized speech?

No experience | Some experience | Experienced | Very Experienced

Appendix D. Introduction

Hello, and welcome to the Human-Computer Interaction Lab. Today, you have the opportunity to participate in our research on how people interact with talking computers.

In this experiment, you will try to find information on certain items in a department store (Hokie Wholesale). The department store has a talking computer database system which provides shoppers with helpful information on store items. Shoppers call the database system on a telephone to find information on selected merchandise. Similarly, you will be using the telephone to find specific information in the database. The talking computer may sound a bit strange at first, but we are sure you will soon be able to understand everything it says. The computer does not understand human speech, but does interpret certain key presses on the telephone keypad as commands.

The database system works by speaking menus of keywords. Keywords are titles for a group of related items (e.g. automotive is a keyword for a group of items like tires, car batteries, and motor oil). When you hear a keyword which most closely relates to the item you are searching for, select that keyword by pressing a defined key on the telephone keypad. The system will then speak a new menu of keywords related to the selected keyword. By selecting the appropriate keywords, you locate the store item in the database. Once you have selected the store item, the computer will speak a short information message about the store item. This message will have something to do with the price, location, availability, or important information about the store item.

Appendix E. 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 dial tone.
2. Press the DIAL key on the telephone keypad (upper right corner).
3. The talking computer will answer the telephone and offer you instructions. Press the **"#" key** and listen carefully to the instruction 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 will indicate when you have completed the experiment. 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 F. Subject's Instructions

The video instructions you just watched included a demonstration of how the telephone information system works and how you should perform the task for this study. The actual telephone information system you will be using today will be similar to the system in the video, but may be different in some ways.

These are the commands that are available to you on the telephone keypad:

To select an item, press the **# key.**

To back-up one menu, press the *** key.**

To select the main menu, press the **0 key.**

When you locate the store item, press the **2 key** to hear the information message.

Appendix G. Database Information Targets and Messages

Message type indicated in parentheses: (I) = Information, (A) = Availability, (P) = Price, and (L) = Location.

1. Target: What is the information message for laundry washers?

   Information message heard: Deluxe models are available with green trimming. (A)

2. Target: What is the information message for football books?

   Information message heard: Faculty discounts are offered to gym teachers. (I)

3. Target: What is the information message for eye mascara?

   Information message heard: Travel supplies are sold for $17.50. (P)

4. Target: What is the information message for men's blazers?

   Information message heard: Garment bags are offered with new purchases. (I)

5. Target: What is the information message for food blenders?

   Information message heard: Boxes and cartons are in the wrapping center. (L)

6. Target: What is the information message for guitars?

   Information message heard: Carrying cases are reduced by 55 to 63%. (P)

7. Target: What is the information message for pearl necklaces?

   Information message heard: Sorority clasps are in the school department. (L)

8. Target: What is the information message for hope chests?

   Information message heard: Walnut stains are reduced by 34 to 40%. (P)

9. Target: What is the information message for silk blouses?

Information message heard: Maternity wear is near ladies lingerie. (L)

10. Target: What is the information message for compact disc recordings?

    Information message heard: Head cleaners are on aisle 12. (L)

11. Target: What is the information message for women's oriental fragrances?

    Information message heard: Manufacturer's samplers are offered to interested shoppers. (I)

12. Target: What is the information message for men's sweaters?

    Information message heard: Rugby letters are sold for $11.60. (P)

13. Target: What is the information message for knit dresses?

    Information message heard: Designer collections are available in red and ivory. (A)

14. Target: What is the information message for gold chains?

    Information message heard: Instant financing is available at the central office. (A)

15. Target: What is the information message for recliner chairs?

    Information message heard: Leather coverings are offered to wholesale buyers. (I)

16. Target: What is the information message for chicken cookbooks?

    Information message heard: Collector editions are available in limited quantities. (A)

Appendix H. Rating Scales

Individual Target Search Ratings

1. Rate how certain you are of your transcription on the following scale:

1	2	3	4	5	6	7
Very Uncertain						Very Certain

2. Rate how difficult it was to understand the information message on the following *scale:*

1	2	3	4	5	6	7
Very Difficult						Very Easy

3. Rate how difficult it was to locate the store item on the following scale:

1	2	3	4	5	6	7
Very Difficult						Very Easy

Post-Experimental Search Ratings

1. Rate the ease of use of the system on the following scale:

1	2	3	4	5	6	7
Very Difficult						Very Easy

2. Rate the intelligibility of the computer voice on the following scale:

1 2 3 4 5 6 7

Very Unintelligible — Very Intelligible

3. Rate the naturalness of the computer voice on the following scale:

1 2 3 4 5 6 7

Very Unnatural — Very Natural

4. Rate how fast the computer talked on the following scale:

Very Slow — Very Fast

5. Rate the speed at which the system responded to your input on the following scale:

1 2 3 4 5 6 7

Very Slow — Very Fast

6. Rate the amount of time you had to respond on the following scale:

1 2 3 4 5 6 7

Very Little — Very Much

7. Rate the menu organization on the following scale:

1 2 3 4 5 6 7

Very Difficult — Very Simple

Appendix I. Subject Debrief

1. Do you like the idea of an information system like this one?
2. Would you use an information system like this one?
3. What applications seem appropriate for an information system such as this one?
4. What improvements would you suggest?
5. Overall, did you like (or enjoy) using this system? :
6. What information would you like to add to the instructions?
7. What would you not include in the instructions?
8. Did you understand the commands?

If not:

   a. Which commands confused you?
   b. What did you understand the command to do?
   c. How did the execution of the command differ from your expectations?

9. Are there any commands you would like to add?
10. Are there any commands you would like to eliminate?
11. What command would you use to restart if you got lost?
12. What command would you use if you wanted to backup one category?
13. Do you think you understand the organization of the data base well enough to use the system comfortably?
14. Did the keyword categories confuse you?
15. What would you change about the experimental session?
16. Was the session length too long?
17. Was the task interesting or boring?

For subjects who heard alternating voices:

18. Did you hear more than one type of voice?
19. Was one more intelligible than the other (which one)?
20. Was one more natural or human sounding than the other?
21. Do you prefer one of these voices over the other?

Appendix J. Performance and Preference Data Summary

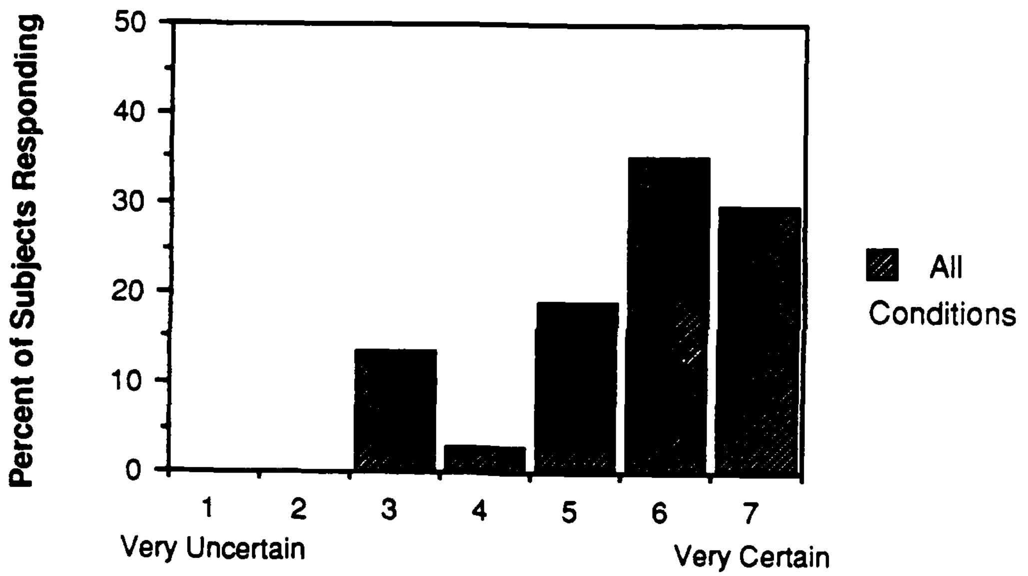

Figure 11. Overall Transcription Certainty Ratings

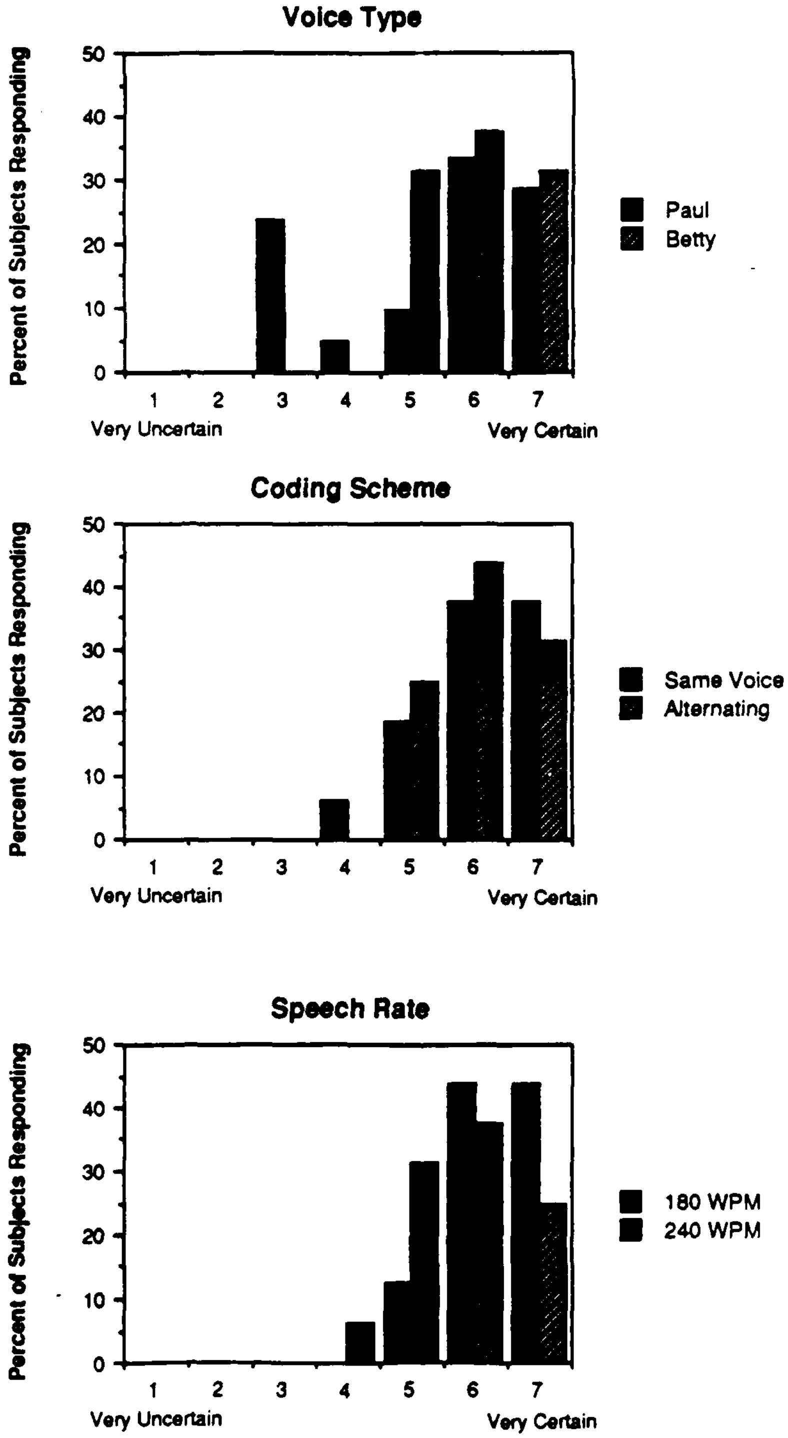

Figure 12. Transcription Certainty Ratings by Voice Type, Coding Scheme and Speech Rate

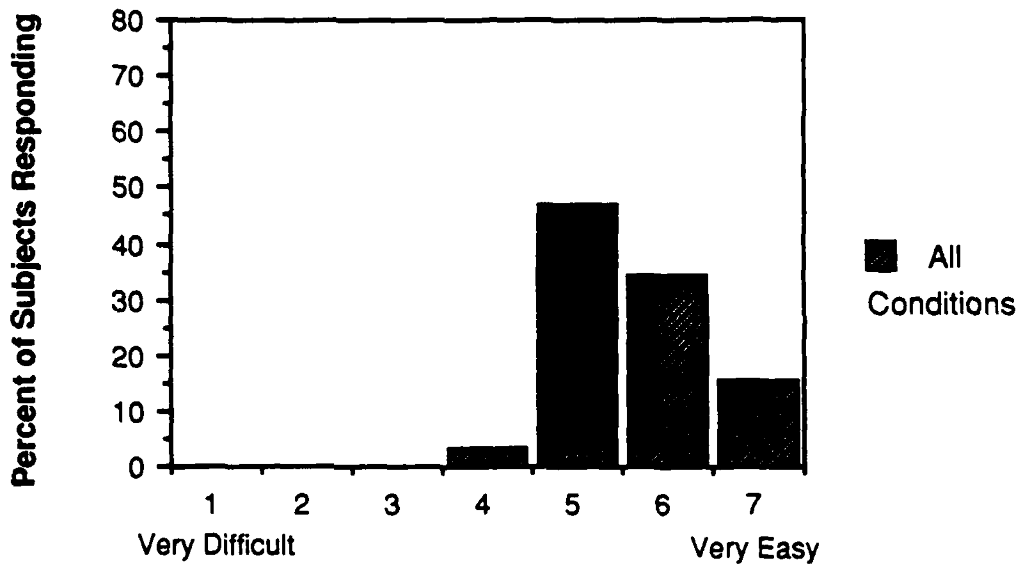

Figure 13. Overall Understanding Difficulty Ratings

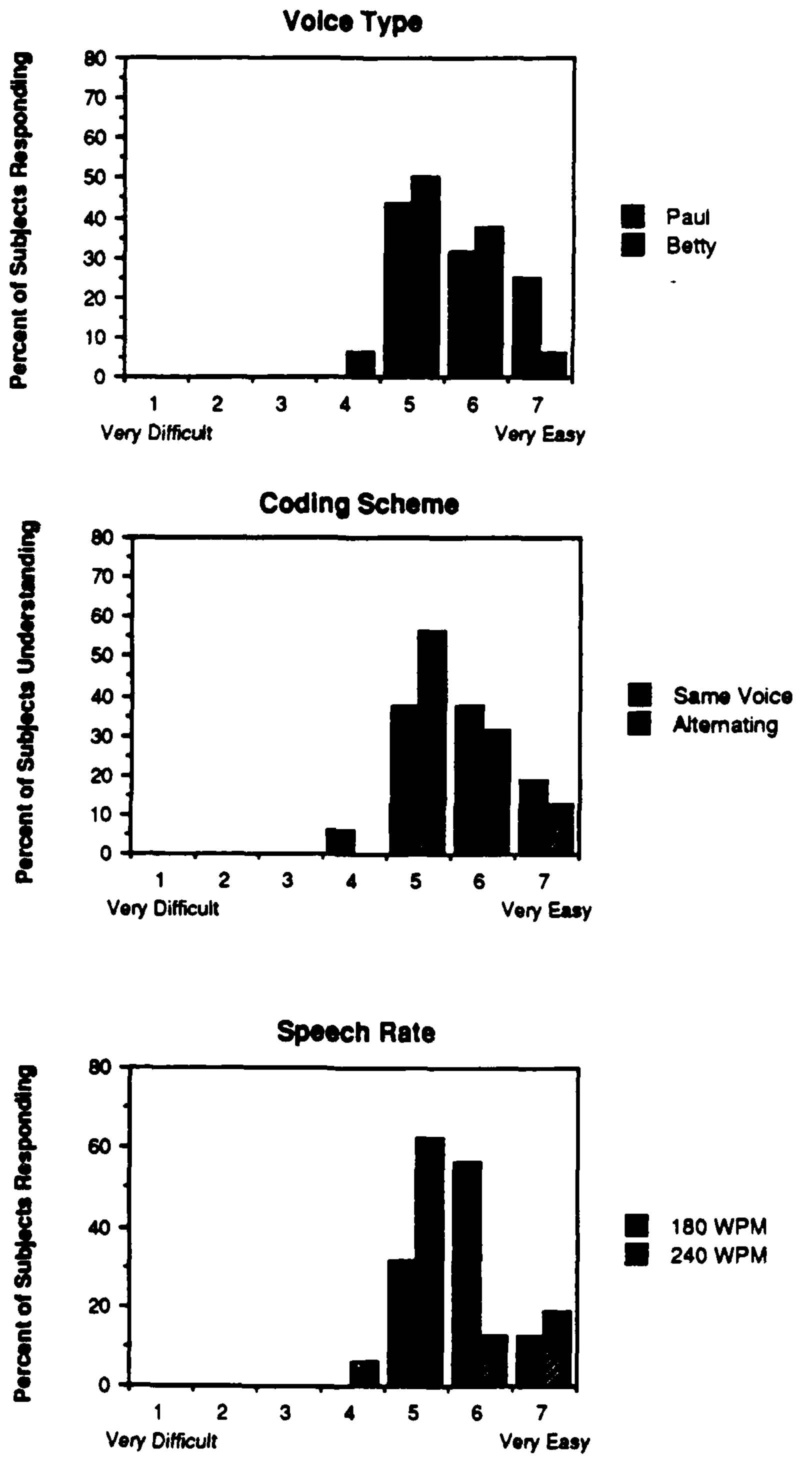

Figure 14. Understanding Difficulty Ratings by Voice Type, Coding Scheme and Speech Rate

Figure 15. Overall Difficulty in Locating Store Item Ratings

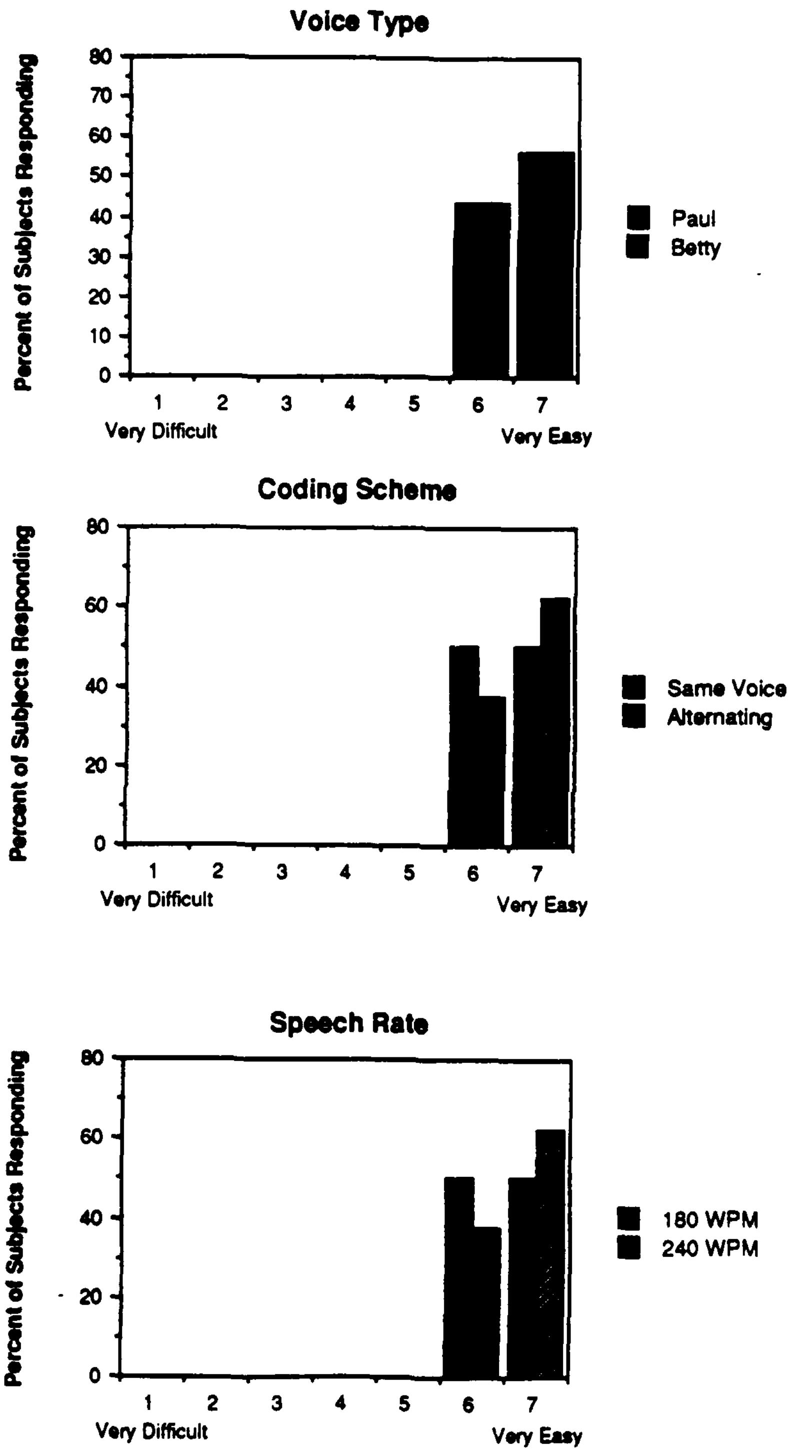

Figure 16. Difficulty in Locating Store Item Ratings by Voice Type, Coding Scheme and Speech Rate

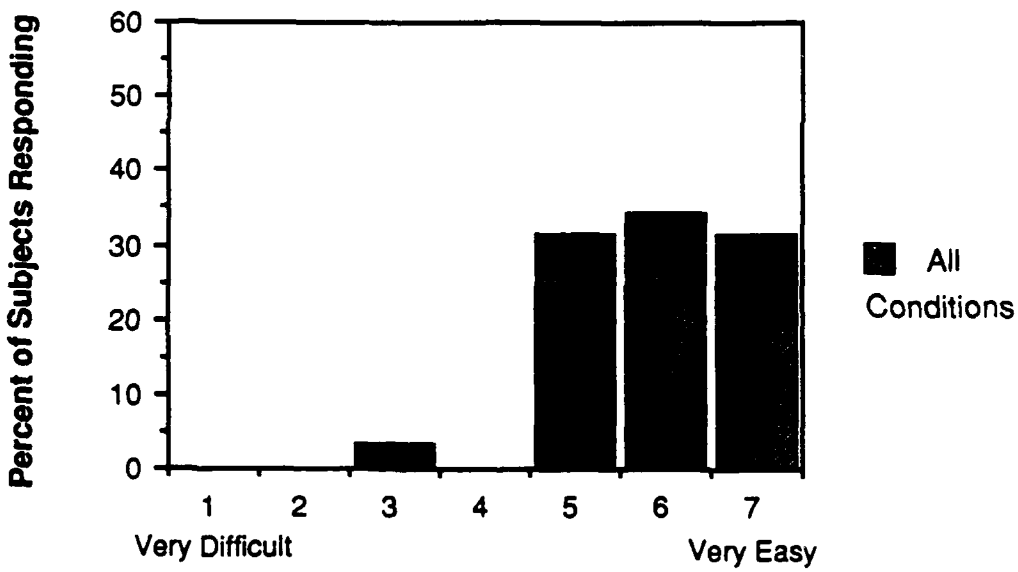

Figure 17. Overall Ease of Use Ratings

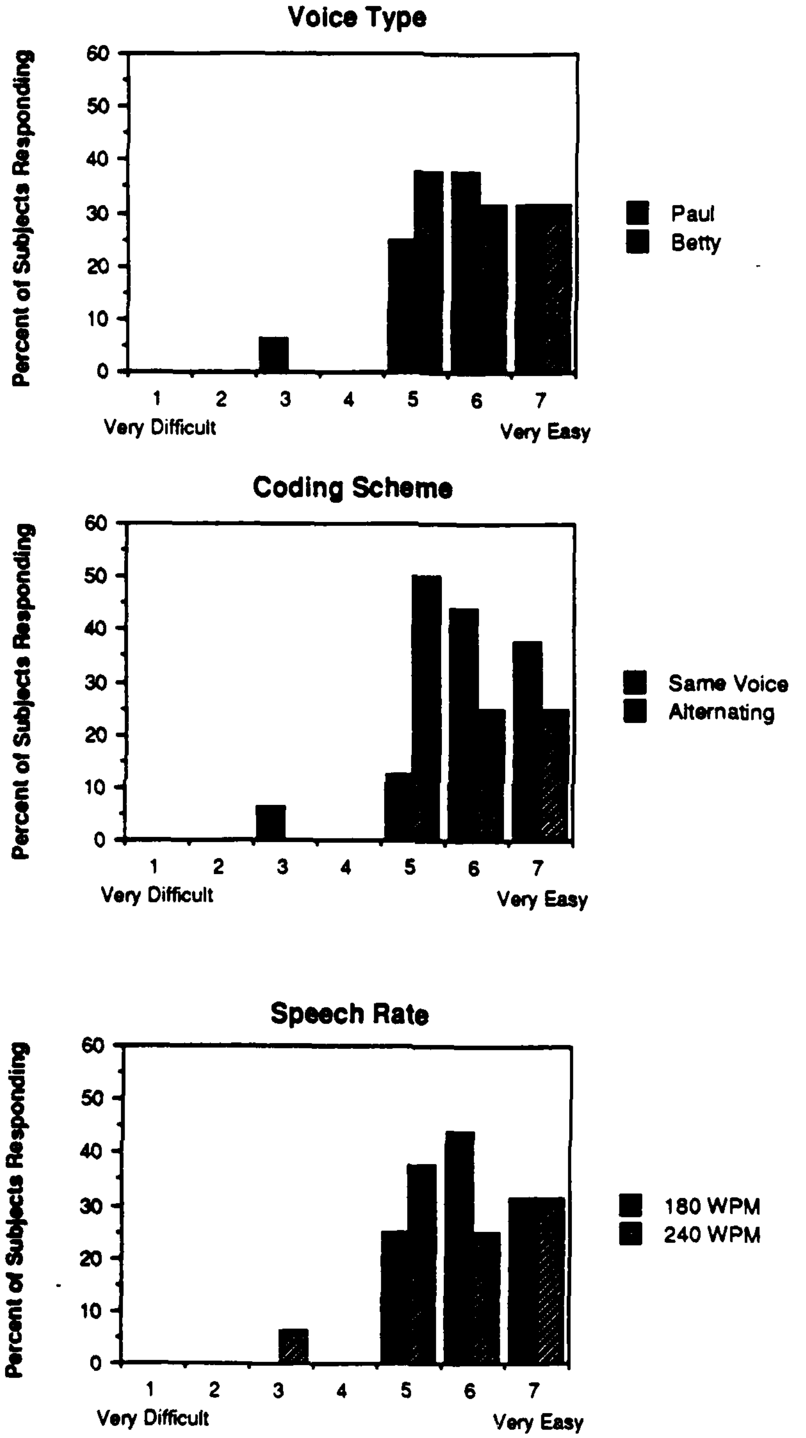

Figure 18. Ease of Use Ratings by Voice Type, Coding Scheme and Speech Rate

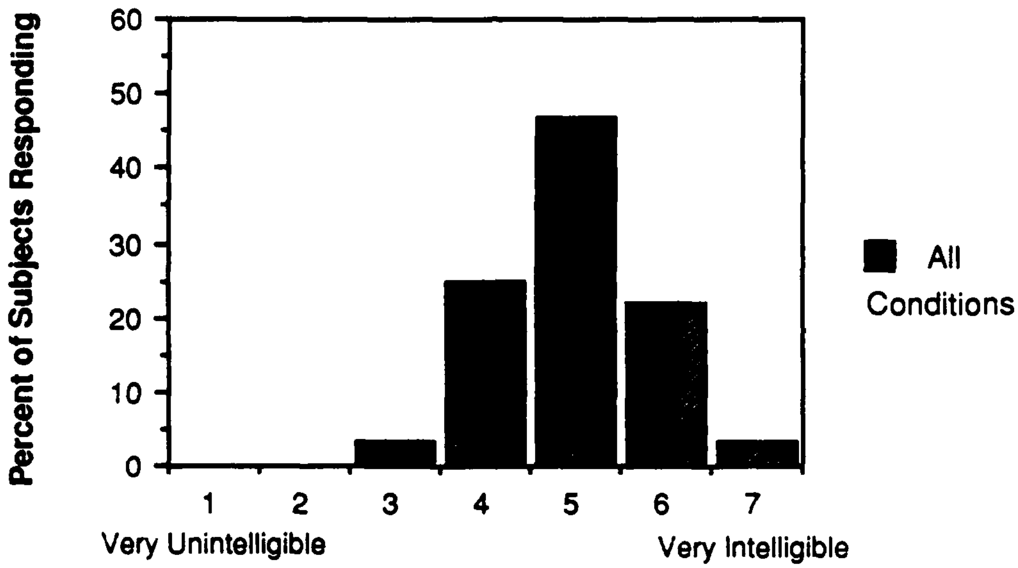

Figure 19. Overall Intelligibility Ratings

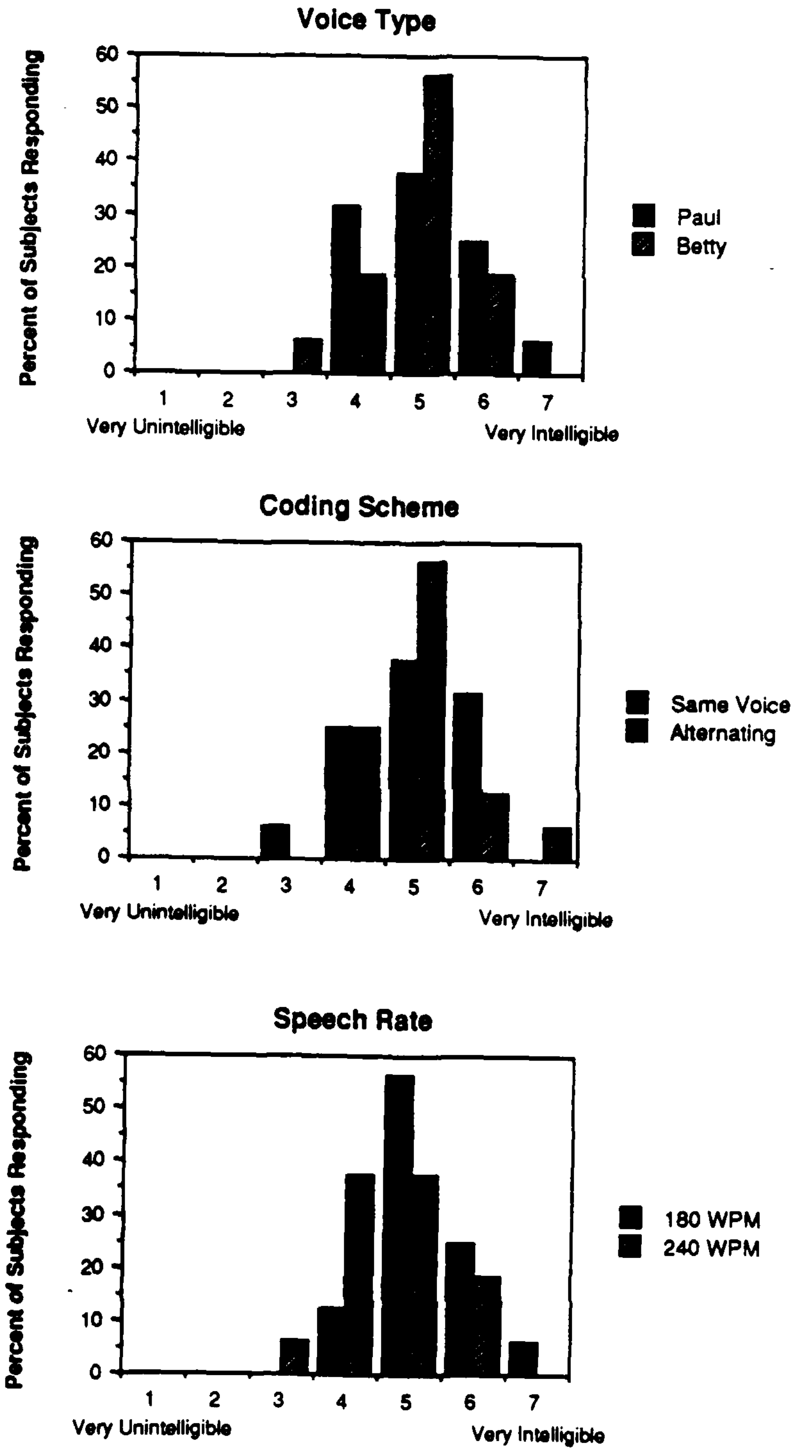

Figure 20. Intelligibility Ratings by Voice Type, Coding Scheme and Speech Rate

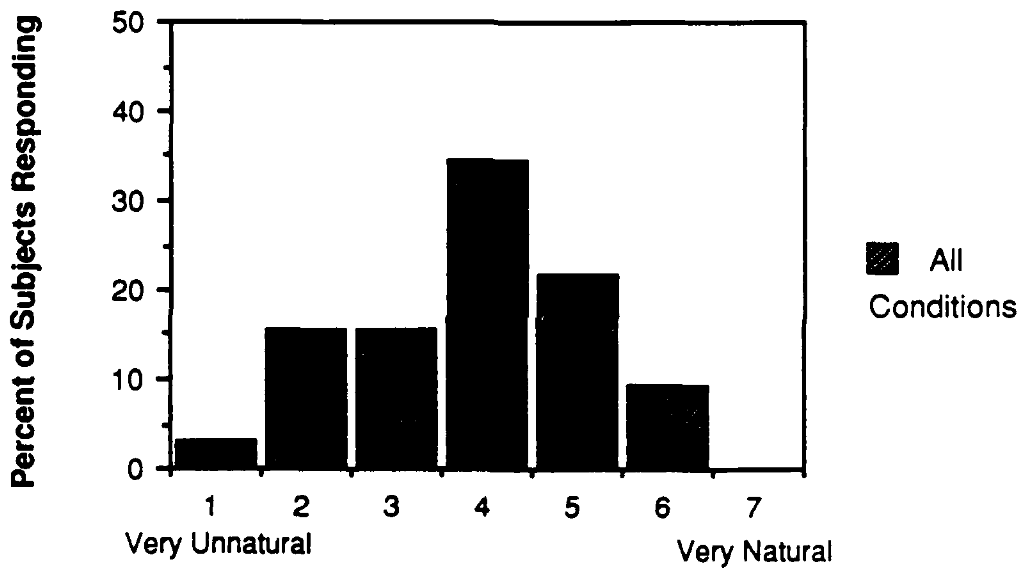

Figure 21. Overall Naturalness Ratings

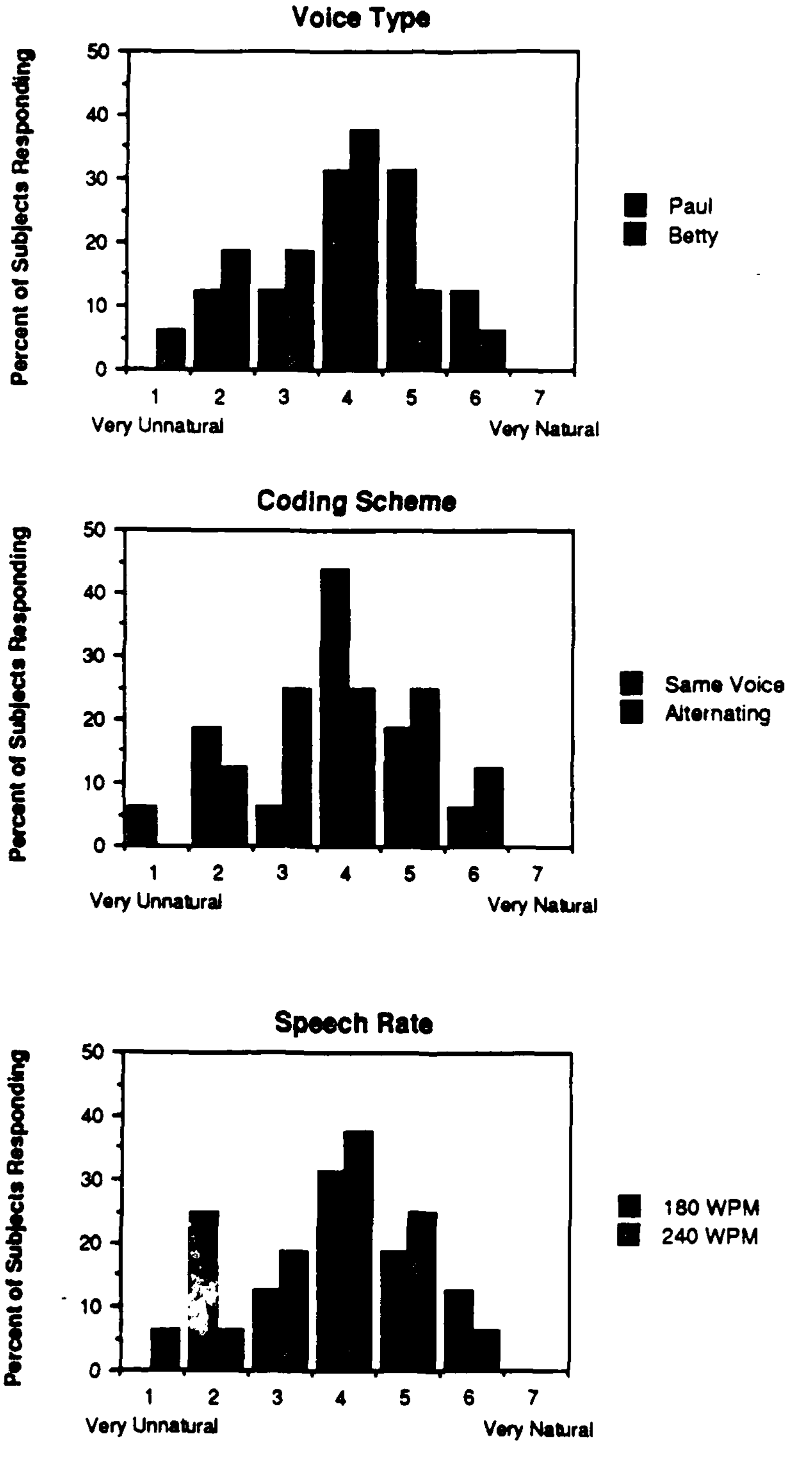

Figure 22. Naturalness Ratings by Voice Type, Coding Scheme and Speech Rate

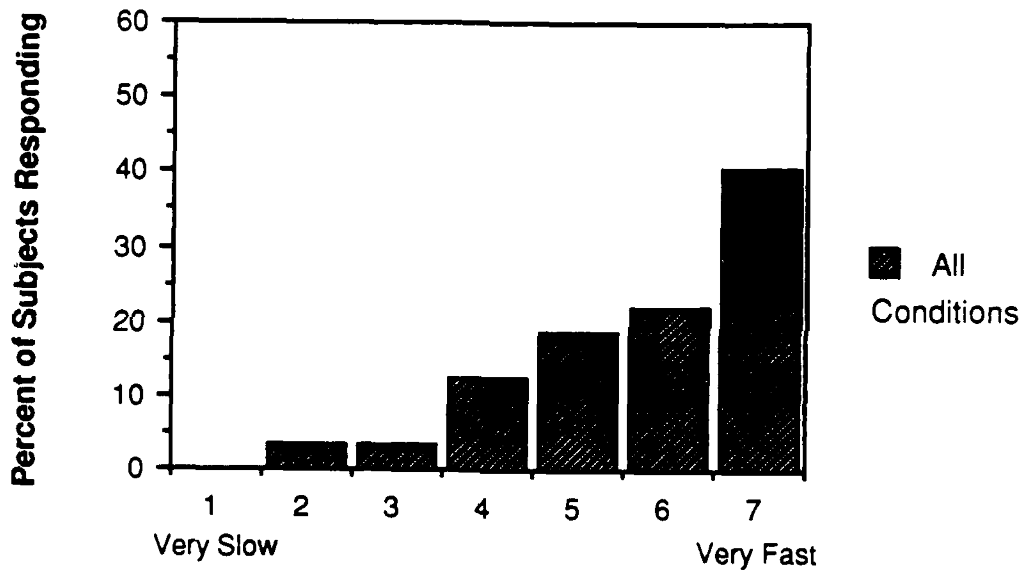

Figure 23. Overall Response Time Ratings

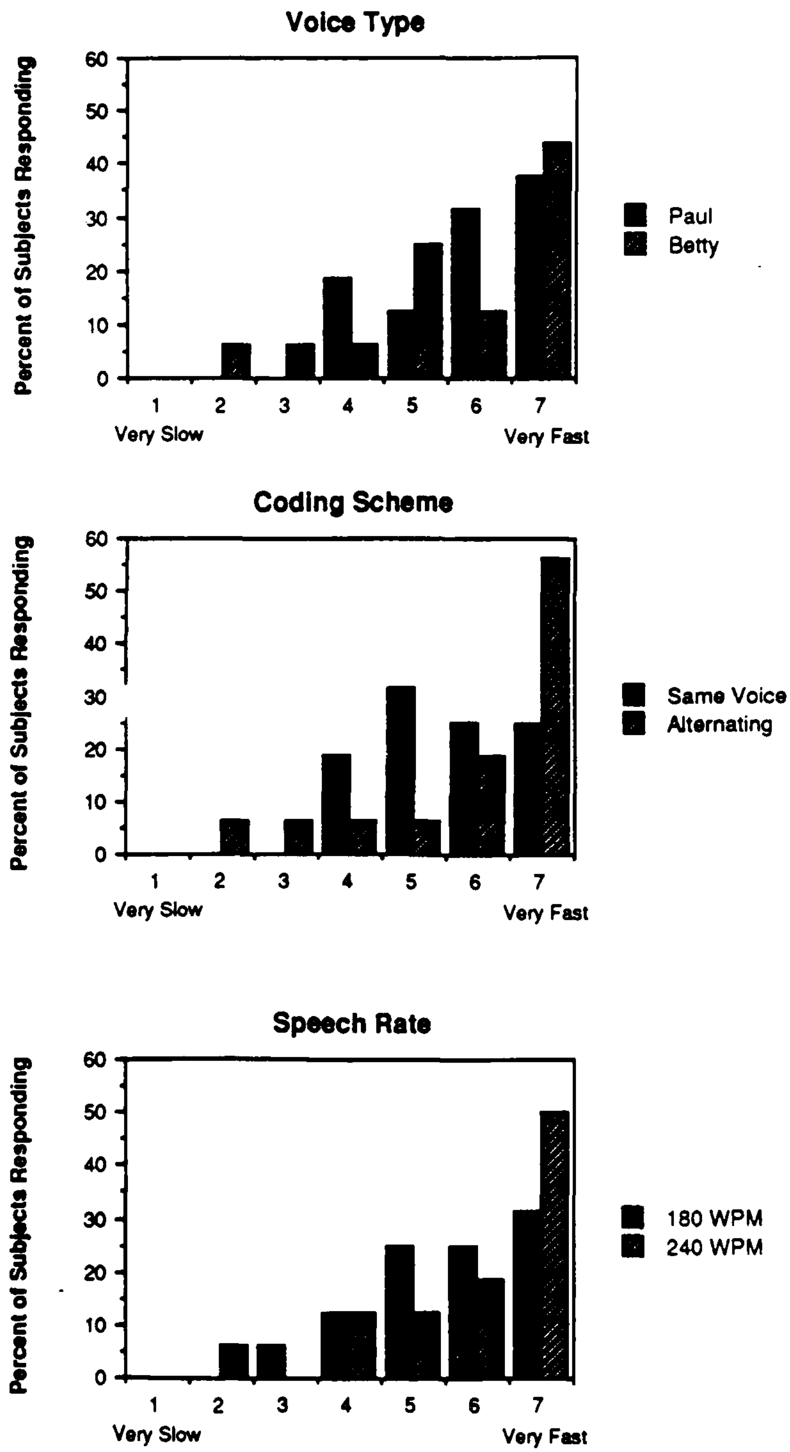

Figure 24. Response Time Ratings by Voice Type, Coding Scheme and Speech Rate

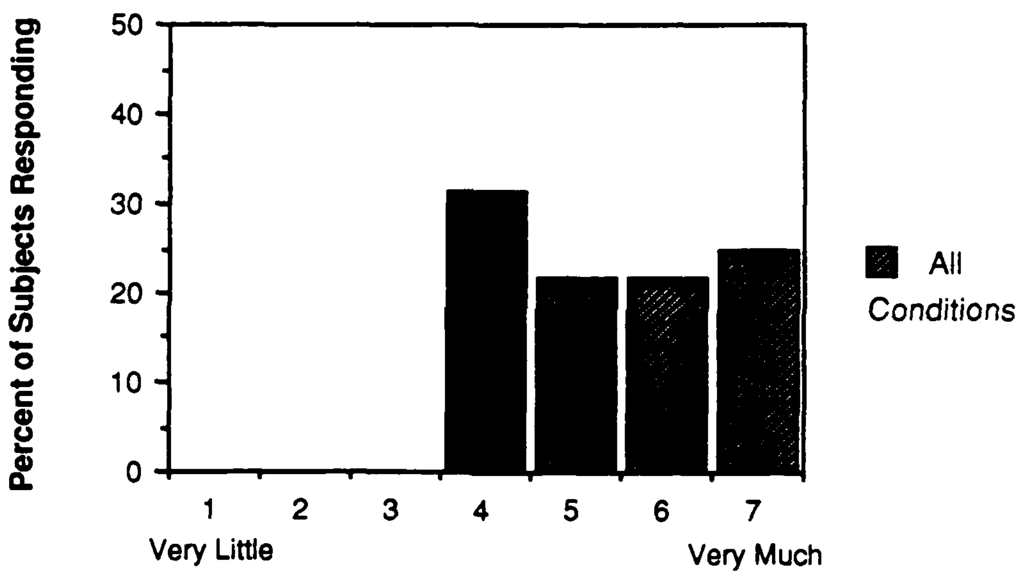

Figure 25. Overall Input Timeout Ratings

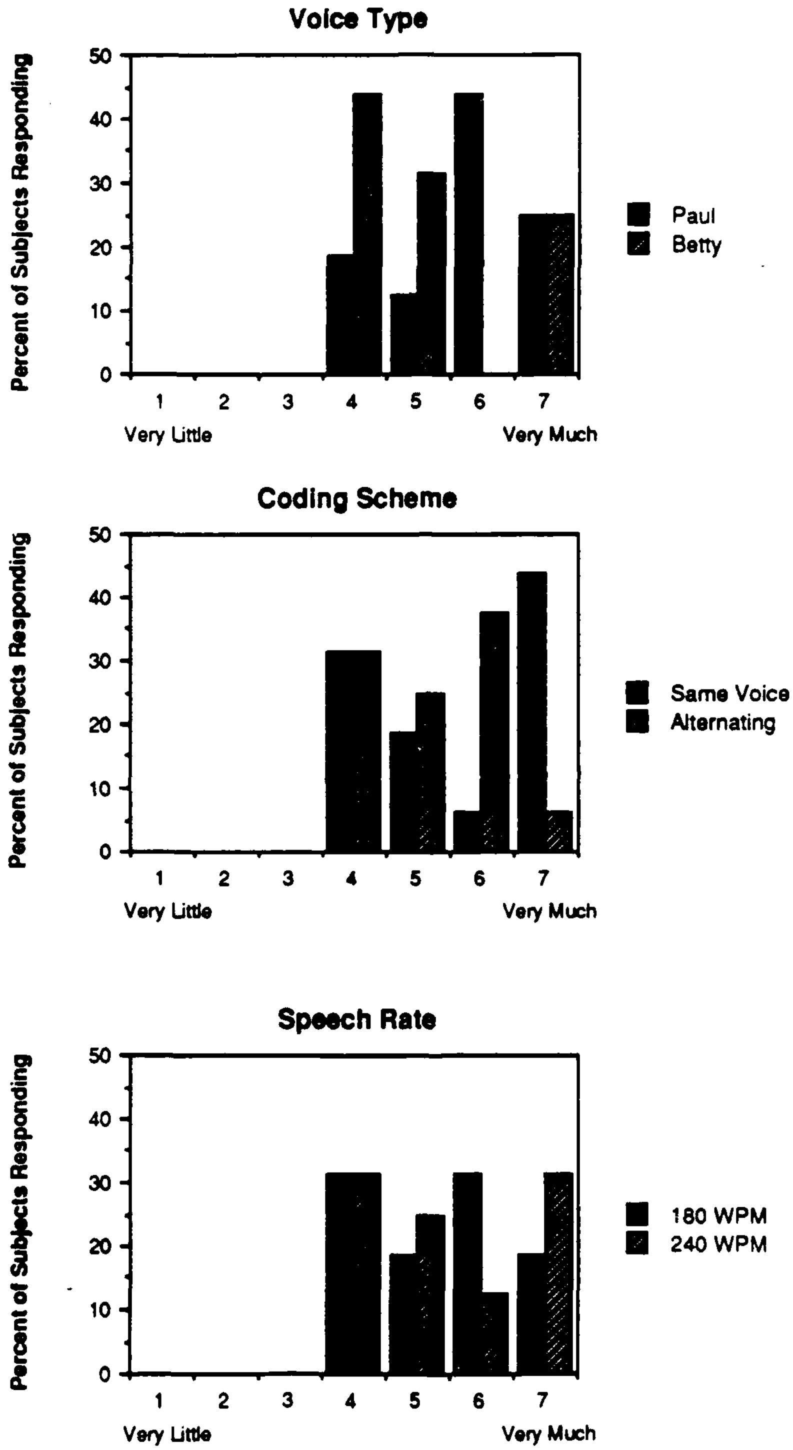

Figure 26. Input Timeout Ratings by Voice Type, Coding Scheme and Speech Rate

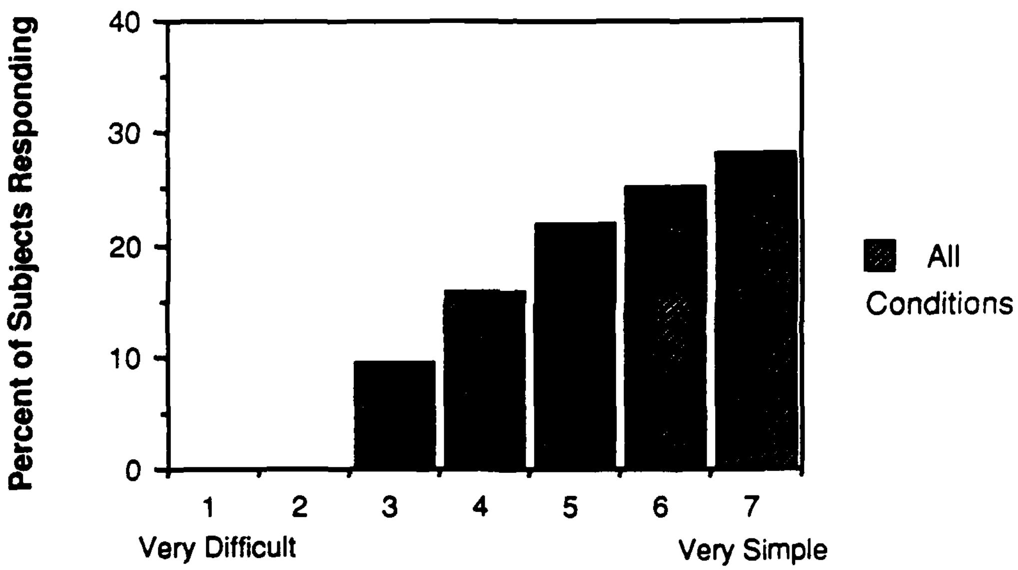

Figure 27. Overall Menu Organization Ratings

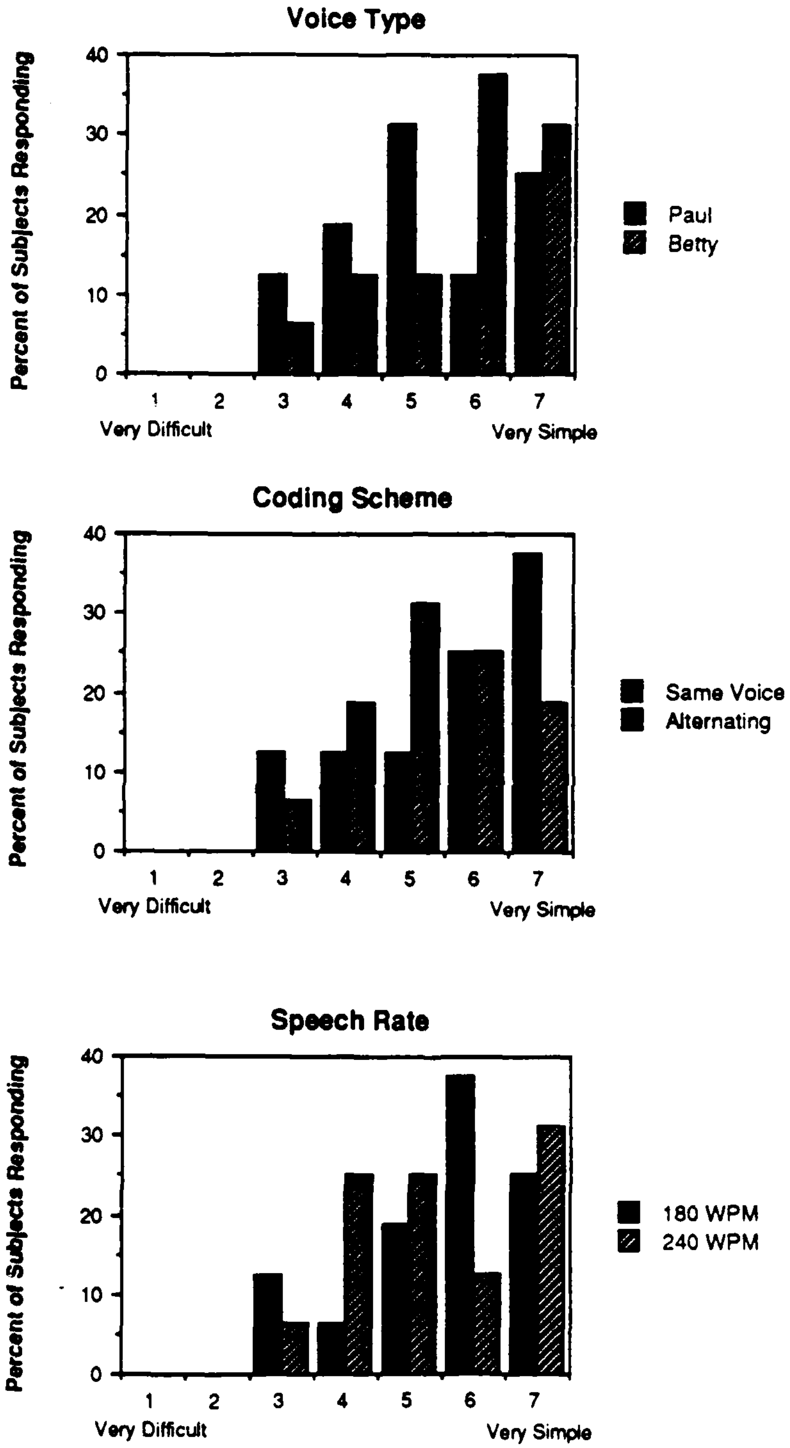

Figure 28. Menu Organization Ratings by Voice Type, Coding Scheme and Speech Rate

	Voice Type	Coding Scheme	Speech Rate	Search Time Ratio	Search Efficiency Ratio	Invalid Keypress Avg.
1	Paul	Same	Slow - 180	.86124	.88889	.0001
2	Paul	Same	Slow - 180	.63908	.73096	.0001
3	Paul	Same	Slow - 180	.57595	.65455	.0625
4	Paul	Same	Slow - 180	.55172	.62882	.0625
5	Paul	Same	Fast - 240	.57115	.69903	.0001
6	Paul	Same	Fast - 240	.67523	.78261	.0001
7	Paul	Same	Fast - 240	.63640	.77838	.0001
8	Paul	Same	Fast - 240	.55633	.63436	.0625
9	Betty	Same	Slow - 180	.82916	.85207	.0001
10	Betty	Same	Slow - 180	.66891	.73096	.0001
11	Betty	Same	Slow - 180	.68572	.73469	.0001
12	Betty	Same	Slow - 180	.62312	.73096	.0001
13	Betty	Same	Fast - 240	.53309	.66667	.0001
14	Betty	Same	Fast - 240	.75889	.87273	.0001
15	Betty	Same	Fast - 240	.74826	.75393	.3125
16	Betty	Same	Fast - 240	.26272	.33333	.0625
17	Paul	Alternating	Slow - 180	.65751	.75393	.0001
18	Paul	Alternating	Slow - 180	.87358	.87805	.0001
19	Paul	Alternating	Slow - 180	.61376	.66359	.0001
20	Paul	Alternating	Slow - 180	.74367	.81818	.0001
21	Paul	Alternating	Fast - 240	.56336	.72000	.0001
22	Paul	Alternating	Fast - 240	.78152	.81818	.1250
23	Paul	Alternating	Fast - 240	.83449	.87805	.0001
24	Paul	Alternating	Fast - 240	.69073	.83237	.0001
25	Betty	Alternating	Slow - 180	.65094	.75789	.0001
26	Betty	Alternating	Slow - 180	.60982	.71642	.0001
27	Betty	Alternating	Slow - 180	.66685	.71287	.1250
28	Betty	Alternating	Slow - 180	.69269	.76596	.1250
29	Betty	Alternating	Fast - 240	.61241	.69231	.0001
30	Betty	Alternating	Fast - 240	.70987	.81818	.0001
31	Betty	Alternating	Fast - 240	.53383	.67606	.0001
32	Betty	Alternating	Fast - 240	.66013	.75393	.0001

	First 8 - Strict	Last 8 - Strict	First 8 - Synonym	Last 8 - Synonym	Total Errors	Total Err (Syn)
1	4	2	4	1	6	5
2	4	3	1	2	7	3
3	10	2	10	1	12	11
4	5	3	5	0	8	5
5	14	6	13	5	20	18
6	5	2	5	2	7	7
7	4	1	4	0	5	4
8	6	2	5	0	8	5
9	2	0	2	0	2	2
10	3	3	2	1	6	3
11	7	4	7	2	11	9
12	2	2	2	0	4	2
13	5	4	4	3	9	7
14	3	6	3	2	9	5
15	7	2	7	0	9	7
16	10	8	10	7	18	17
17	6	2	6	1	8	7
18	3	4	3	1	7	4
19	4	4	4	3	8	7
20	5	2	5	0	7	5
21	10	4	8	0	14	8
22	6	1	6	1	7	7
23	10	2	10	1	12	11
24	5	5	5	3	10	8
25	4	3	3	1	7	4
26	1	1	1	0	2	1
27	1	3	1	1	4	2
28	5	2	5	1	7	6
29	10	6	10	3	16	13
30	7	4	5	3	11	8
31	7	5	6	4	12	10
32	2	4	2	0	6	2

	Tran/Cert	Diff/Under	Locat/Diff	Ease/Use	Intelligibility	Naturalness
1	7	7	7	7	6	5
2	5	5	6	6	5	4
3	6	6	6	6	6	5
4	7	7	7	7	6	6
5	7	7	7	6	6	4
6	6	6	7	6	4	4
7	4	5	6	3	4	2
8	7	6	6	7	4	4
9	6	6	7	7	6	4
10	7	6	6	7	5	2
11	7	5	6	6	5	2
12	6	6	7	6	5	5
13	5	5	6	5	3	4
14	6	5	7	6	5	4
15	6	5	6	7	5	3
16	5	4	7	5	4	1
17	6	6	7	6	7	6
18	6	5	7	5	5	4
19	7	5	7	7	5	3
20	7	6	6	5	4	2
21	5	5	6	5	5	3
22	6	5	6	5	5	5
23	6	7	7	7	5	5
24	6	5	7	6	4	5
25	5	6	6	5	5	3
26	6	6	6	6	5	4
27	7	6	7	6	5	4
28	6	5	6	5	4	2
29	5	5	7	5	5	4
30	5	5	7	5	4	3
31	7	7	7	7	6	5
32	7	5	7	7	6	6

	Speech/Rate	Response Time	Input Timeout	Menu Organization
1	5	6	6	7
2	4	5	4	3
3	4	4	7	6
4	4	6	4	7
5	6	4	7	5
6	5	6	7	7
7	4	5	5	4
8	6	7	7	7
9	4	5	7	7
10	5	7	4	6
11	3	6	5	6
12	5	5	7	6
13	4	4	5	5
14	6	7	7	7
15	6	7	4	4
16	6	5	4	3
17	4	7	6	5
18	3	4	6	6
19	4	7	6	5
20	4	7	6	3
21	6	6	6	4
22	6	7	4	5
23	5	6	6	4
24	7	7	5	5
25	5	3	4	4
26	5	7	5	5
27	5	5	4	7
28	3	6	5	6
29	6	7	4	6
30	7	2	5	7
31	7	7	7	7
32	5	7	4	6

Vita

DAVID WOOD HERLONG

PERSONAL

Date of Birth: June 8, 1952

Place of Birth: Montgomery, Alabama

PRESENT STATUS

Major
United States Air Force
Instructor, Department of Behavioral Science and Leadership (DFBL)
United States Air Force Academy, Colorado

EDUCATION

Master of Science, Industrial Engineering - Human Factors, Virginia Polytechnic Institute and State University, Blacksburg, Virginia. September 1986 to May 1988.

Bachelor of Science, Clinical Psychology, United States Air Force Academy, Colorado. July 1971 to June 1975.

PROFESSIONAL EXPERIENCE

1985 - 1986 F-16A/B Flight Examiner and Assistant Chief, Wing Standardization and Evaluation, 8th Tactical Fighter Wing, Kunsan Air Base, Republic of Korea.

1986 - 1985 Flight Commander and F-16A/B Flight Leader, 16th Tactical Fighter Squadron, 388th Tactical Fighter Wing, Hill AFB, Utah.

1981 - 1982 OV-10A Replacement Training Unit (RTU) Flight Instructor Pilot, 549th Tactical Air Support Training Squadron, 549th Tactical Air Support Training Group, Patrick AFB, Florida.

1981 - 1979 OV-10 Squadron Instructor Pilot, 20th Tactical Air Support Squadron, Sembach Air Base, Federal Republic of Germany.

1977 - 1979 A-7D Aircraft Commander, 75th Tactical Fighter Squadron, 23rd Tactical Fighter Wing, England AFB, Louisiana.

1975 - 1976 United States Air Force Flight Training, Craig AFB, Alabama.

HONORS

United States Air Force Meritorious Service Medal with 1 Oak Leak Cluster
United States Air Force Commendation Medal with 1 Oak Leak Cluster
United States Air Force Achievement Medal
United States Air Force Combat Readiness Medal with 1 Oak Leak Cluster
United States National Defense Service Medal
Distinguished Graduate, Squadron Officer School, Maxwell AFB, Alabama.
Alpha Pi Mu, Industrial Engineering Honor Society
Phi Kappa Phi, National Scholastic Honor Society

David W. Herlong