The Effects of Recognition Accuracy and Vocabulary Size Of A Speech Recognition System on Task Performance and User Acceptance

by

Sherry P. Casali

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 D. Dryden

Robert D. Dryden, Chairman

Beverly H Williges

Beverly H. Williges

Paul T Kemmerling

Paul T. Kemmerling

May 13, 1988

Blacksburg, Virginia

The Effects of Recognition Accuracy and Vocabulary Size Of A Speech Recognition System on Task Performance and User Acceptance

by

Sherry P. Casali

Robert D. Dryden, Chairman

Industrial Engineering and Operations Research

(ABSTRACT)

Automatic speech recognition systems have at last advanced to the state that they are now a feasible alternative for human-machine communication in selected applications. As such, research efforts are now beginning to focus on characteristics of the human, the recognition device, and the interface which optimize the system performance, rather than the previous trend of determining factors affecting recognizer performance alone. This study investigated two characteristics of the recognition device, the accuracy level at which it recognizes speech, and the vocabulary size of the recognizer as a percent of task vocabulary size to determine their effects on system per'ormance. In addition, the study considered one characteristic of the user, age. Briefly, subjects performed a data entry task under each of the treatment conditions. Task completion time and the number of errors remaining at the end of each session were recorded. After each session, subjects rated the recognition device used as to its acceptability for the task.

The accuracy level at which the recognizer was performing significantly influenced the task completion time as well as the user's acceptability ratings, but had only a small effect on the number of errors left uncorrected. The available vocabulary size

also significantly affected the task completion time; however its effect on the final error rate and on the acceptability ratings was negligible. The age of the subject was also found to influence both objective and subjective measures. Older subjects in general required longer times to complete the tasks; however, they consistently rated the speech input systems more favorably than the younger subjects.

Acknowledgements

The author wishes to express appreciation to Dr. Robert D. Dryden for the support and guidance he provided as committee chairman throughout the course of this research. Special thanks are also due to committee members Beverly H. Williges for the direction and expertise she provided, and Professor Paul T. Kemmerling for his advice and encouragement. I also wish to thank Calvin L. Selig for the development of the speech recognition system simulation software.

On a more personal note, I would like to thank my parents for a lifetime of support and encouragement. And finally, I would like to thank my husband John, whose patience and encouragement have helped make the past two years an enjoyable as well as rewarding experience.

Table of Contents

List of Illustrations

List of Tables

Introduction

Automatic speech recognition refers to the ability of a machine to discriminate spoken utterances. The first such devices were developed in the mid 1950's, and significant progress has been made within the past decade. It has long been known that unconstrained speech is the fastest, most efficient means for two people to communicate (Chapanis, 1975). Now, applying spoken communication to human-machine systems is proving to be an efficient alternative for many applications. Such applications include baggage and postal sorting, quality control on production lines, and voice direction of machine tools. Voice input has the advantage that no typing skills are needed and that it frees the hands and eyes to perform concurrent tasks.

One application which appears particularly suited for speech recognition is as computer input devices for the physically disabled. Conventional computer input devices may be useless to a disabled user, and the specially designed aids in use today are often slow and cumbersome. Speech recognition allows disabled users to perform many tasks quickly and effortlessly which otherwise are difficult or even impossible. Voice controlled wheelchairs and environmental control systems have been developed and are being used in a limited number of applications (Cohen and Graupe, 1980; Damper, 1984; Youden, Sell, Reich, Clagnaz, Louie, and Kolwicz; 1980). Speech input text-processing aids are being developed to allow disabled persons to compose letters and reports quickly (Damper, 1984). Glenn, Miller, and Broman (1976) report success in using voice input to process text as well as program a computer. These developments may open a wide range of employment opportunities for disabled persons in industrial and office environments.

Current speech recognition systems, however, suffer several limitations. This study examined perhaps the two most critical of these limitations. First, currently available low-cost recognition devices frequently make mistakes in discriminating what the user says. And secondly, most currently available systems are only able to recognize a limited number of words. If speech recognition devices are going to be used, whether by able-bodied or disabled individuals, the effects of these limitations on the overall performance of the system should be known.

Purpose

The purpose of this study was to investigate the degree to which recognizer accuracy and vocabulary size affect measures of system performance. These measures included the objective measures of time to complete the data entry task and number of errors remaining in the data at the end of the task as well as subjective measures of user acceptability. In addition, different age groups of subjects were used in order to determine if the age of a user plays a role in his or her ability to use a speech recognition system successfully and its effects on subjective assessment of speech as an input mode.

Background and Literature Review

Overview of Speech Technology

A speech recognition system is a device which takes spoken input from a human user and converts it to a language which is understandable by a computer. Many currently available low-cost systems rely on a template matching method for recognizing incoming speech, i.e., they compare each utterance with previously stored speech patterns. A recognizer is initially "trained" by having one or more persons speak each word of the allowable vocabulary one or more times. For each utterance, the recognizer first determines the beginning and end points and then converts the acoustical waveform into an analog electrical signal. This signal is then filtered to extract the information-carrying features, which are then converted into a digital pattern. This pattern is combined with other patterns of the same word forming a template which is then stored in memory. During the recognition phase, each input item is processed in the same manner as during training. The resulting pattern is then compared with all the stored reference templates to determine which pattern in the most similar. Given that the most similar reference template scores above some predetermined acceptance level, the system accepts the reference template as being the incoming utterance. This may result in the spoken word being displayed on a CRT terminal, some command being executed, or any number of other responses just as if the word had been entered thru more conventional input devices (e.g. keyboard, mouse, etc.). A more detailed explanation of the recognition process may be found in Lea (1980) or Reddy (1976).

Speech recognition devices are usually classified as recognizing either discrete, connected, or continuous speech. Discrete speech recognizers require that the user separate utterance with a brief pause, typically 50 to 200 milliseconds. A single utterance may consist of a single word or a short string of connected input. Connected speech recognizers do not require the intermediate pause between inputs, but are able to detect word boundaries within a string of connected speech. They do, however, require that the user carefully annunciate each word. Continuous speech recognition systems allow the user to speak in a natural rhythm.

Discrete, connected, and continuous speech recognition systems can be classified further as either speaker-dependent or speaker-independent systems. Speaker-dependent systems require that each user enter several samples of each word in the vocabulary to form the reference templates used during recognition. In other words, his voice patterns are compared only with prerecorded samples of his own speech. Speaker-independent systems do not require each user to enter voice samples first, rather they compare a user's input with previously collected samples from a number of other speakers.

Emphasis on System Performance

Speech recognition is a new technology. Although the first machine that could recognize speech was developed thirty five years ago (Davis, Biddulph, and Balashek; 1952) the field is often described as still being in an "infancy stage" (Lea, 1980) because progress has been slow and difficult. It has only been within the past decade that recognizers have progressed to the state that they are being used in actual applications and being considered for many more. In the past, the primary measure of recognizer performance has been the accuracy with which spoken words were recognized. If a new recognizer being developed were able to recognize correctly more words than another system, it was considered better. This performance measure is appropriate for measuring the relative accuracy of different recognition algorithms. But now that speech recognition systems are finding their way into actual applications, a measure of overall system performance is also necessary.

When a speech recognition device is incorporated into a human-machine system, the ultimate performance measure is of course the effectiveness of the entire system. This measure is dependent on the individual task. It may be the total time to complete the task, the number of errors resulting, the reduction in total costs for task completion, user satisfaction, or a combination of these and other measures. This approach of evaluating recognition systems has gained increasing interest (e.g. Grady, 1982; Harris, 1982; Schmandt, 1982; Simpson, McCauley, Roland, Ruth, and Williges; 1985) but because of its newness, has yet to be employed in many research efforts.

As such, one of the first steps in a systems approach of evaluating speech recognition devices is to identify factors associated with the use of speech recognition which affect the overall performance of the system. These factors may include characteristics of the user, the recognition device, and the interface between the two. This research will investigate two characteristics of the speech recognizer, accuracy level and vocabulary size, and one user characteristic, age.

Accuracy

The accuracy of a speech recognition system refers to the percentage of time that the recognizer correctly identifies an input utterance. Unidentified inputs (recognition errors) can be of several types: substitution, deletion, rejection, and insertion (Baker, 1982). A substitution error, also referred to as a misrecognition, occurs when the system incorrectly identifies the input utterance as some other word in its allowable vocabulary. The user may speak the word "nine" but the recognizer incorrectly chooses the word "five". A deletion error is the result of the system failing to respond to a valid input. This type of error would occur, for example, if the amplitude of a spoken input is not great enough to elicit a response from the system. Rejection errors occur when the system correctly identifies the input but the template's "closeness" score is below the acceptable level or the second best score is unacceptably close to the best match resulting in a decision by the speech recognition algorithm to reject the input. Because the results are the same for the user, often no distinction is made between deletion and rejection errors, and together they are referred to as nonrecognition errors. Finally, an insertion error occurs when the system incorrectly identifies some sound outside the acceptable vocabulary (e.g. a cough, background noise, or a nonvocabulary word) as a vocabulary item. When manufacturers test recognition devices, the tests are usually carefully controlled to eliminate any such spurious noises and therefore, the most commonly reported errors are misrecognition and nonrecognition errors.

Difficulties in Obtaining High Accuracy

Because accurate recognition is dependent on matching an incoming speech signal with a stored signal, any factor which alters this signal, either at the source (speaker), the receiver (machine), or in-between has the potential of affecting the recognition accuracy. As a result, high recognition levels are often difficult and expensive, if not impossible to obtain. Manufacturers usually claim accuracy levels of 98-99%, however, these results are obtained under controlled laboratory conditions. When the system is placed in an operational environment where the speech signal is subject to influencing factors, recognition errors much greater then 1-2% can be expected. The fact that the manufacturer achieved such high accuracy during testing says little about the system's ability to recognize speech in normal operational conditions (McCauley, 1984).

Lea (1982) reports a structured list of over 80 variables which he suspects as influencing recognition accuracy. Some of these variables have been observed as influencing accuracy throughout the development of speech recognizers. Others have been the subject of empirical investigations focused specifically on determining their influence on recognition accuracy. Still others are merely suspected of influencing accuracy. Table 1 lists these variables. As the list indicates, these factors include characteristics of the user (age, sex, experience, workload, etc.), the language (vocabulary size, word length, word confusability, etc.), the channel and environment (noise, vibration, microphone type and placement, etc.), and the recognition algorithm (sampling process, word boundary detection, etc.) among many others.

A thorough review of all of these factors would not be worthwhile for the purposes of this thesis, however, a few selected factors will be discussed with respect to their influence on recognition accuracy in order to illustrate the difficulties and trade-offs involved in obtaining high accuracy levels.

The ultimate goal of speech recognition developers is to design a system which can understand speech from any number of users speaking in a natural rhythm. This goal has only been reached to a limited degree. Recognizers have been developed which recognize continuous speech from a number of different speakers, however, they are not in practical use because of accuracy limitations.

As discussed previously, speaker-independent systems do not require a user to train the system with his own voice patterns before use, but instead compare each input with a large number of stored samples from different individuals. Developers noted early that because inter-speaker variability is so much greater than intra-speaker variability, speaker-dependent recognizers can achieve higher recognition rates than speaker-independent systems (Doddington and Shalk, 1981). The more varied the group of speakers using the system, the more difficult the recognition process. Rollins (1984) notes that it is difficult for a speaker-independent system to recognize accurately both male and female speakers. When a recognizer was trained by four male speakers and then tested by each of them, the median recognition accuracy

Table 1. Factors Affecting Error Rates. From Lea (1982)

Human Factors

- form of speech (isolated, connected continuous)
- rate of speaking
- speaker dependent or independent
- sex of speaker
- speaker's vocal tract size
- speaker dialect
- speaker's habits of pronunciation
- speaker's glottal spectrum
- physical state of the speaker
- simultaneous workload and distractions
- psychological state & stress on speaker
- speaker experience & skill
- motivation of the speaker
- training method
- number of training samples / word
- training sequence
- time of day
- time of week
- time since training

Channel and Environmental Factors

- noise level
- type of noise
- communication bandwith
- spectral distortion
- vibration forces
- acceleration on speaker
- microphone type
- microphone placement
- signal amplitude
- tape recording distortions
- time stretching on analog tape, where analog tape is is used for testing w/ a database
- temperature
- humidity

Language Factors

- size of active subvocabulary
- length of word
- number of syllables
- language spoken
- manner, place & voicing of the initial final, and medial consonants
- vowel pattern in word
- degree of coarticulation
- enhanceability of the vocabulary so that the device can handle new words
- stress in the word
- intonation in word
- rhythm patterns of prosody in word

Task Factors

- type of device
- manufacturer & model
- interword pause duration
- interface w/ system
- prior use & handling of device
- manufacturer's field experience w/ real uses of recognisers

Performance and Response Factors

- type of error
- verification of decisions by visual (or other) display
- feedback of intermediate results
- procedure for correcting errors

Algorithmic Factors

- alignment of discret sampling w/ waveform
- amplitude & time normalization procedures
- hardware errors & variabilities in hardware characteristics among recognisers of same model
- positioning of word boundary locations
- resolution in the frequency spectrum
- set of acoustic parameters extracted
- degree of focus on speech distinguishing features
- number of time segments into which the utterance is sliced
- local distance measure for comparing short time segments & received acoustic pattern
- method of combining local distances into full-word distance for pattern matching
- pattern matching control strategy
- bits/word in reference pattern
- rejection threshold & method of rejecting errors that are confusably similar to more than 1 allowable word
- degree of dependence of algorithm on invariance of acoustic distortion in pattern
- degree of context-dependency or non-linearity of segment sequences in expected
- allowed syntactic complexity
- type of grammar used to characterise allowed structural combinations
- semantic & pragmatic variety permitted in discourse w/ the machine

obtained was as high as 97.6%. However, when one of the four speakers was female and the members trained and tested the device, the median accuracy declined to below 90%. Similarly, different geographical dialects, ages, etc. within a user population may result in even lower recognition accuracy rates for speaker-independent systems.

Most currently available speech recognition systems are of the isolated-word variety, requiring a distinct pause between words. This is because connected and continuous recognition systems are still much less accurate than isolated-word systems (Reddy and Zue, 1983). As the rate of speech increases, word boundaries become shorter, or even nonexistent making it difficult to determine where one word ends and the next begins. For example, in the phrase "up and down", the ending "d" in the word "and" is often merged with the beginning sound of the word "down", usually without even a brief pause between the words. Additionally, not only is the duration of a word spoken in succession with other words shorter, but the shortening is highly nonlinear. Generally, vowel sounds are shortened substantially more than consonants, thus changing the pronunciation of the word making template matching more difficult (Rabiner, Bergh, Wilpon, 1982). Also, the pronunciation of each word is influenced by the words preceding and following it. This effect is known as coarticulation and may significantly alter the speech signal of words spoken in succession (Klatt, 1980). For exampla, the "d" and "y" in "Did you ..." often become a "j" sound when spoken rapidly. As a result, high accuracy is more easily obtained with isolated speech systems than continuous speech systems and will be for some time.

Perhaps the most investigated environmental factor affecting recognizer accuracy is the ambient noise level (e.g. Coler, 1982; Drennen, 1980; Kersteen, 1982; Rollins and

Wieson, 1983; Simpson, Coler and Huff, 1982). High ambient noise levels can affect recognition in two ways. Most obviously, the background noise can mask the speech signal reducing the signal to noise ratio. In addition, under high noise levels speakers tend to change their speech. In particular, speaking rates decrease, intensity and duration of utterances increase, vocal pitch increases and even the acoustic-phonetic structure of the speech is altered (Pisoni et. al., 1984). Though further research is needed to understand thoroughly the effects of different types and levels of noise on a recognition system, the general agreement is that recognition in low ambient noise conditions is usually only slightly more accurate than in high noise environments provided the system used in high noise conditions is also trained in the high noise conditions. The real problem arises when the recognizer is trained in one noise environment and used in another. Kersteen's study (1982) reports as much as an 18% decrease in accuracy when the system is trained in a quiet environment and used under higher noise levels. As a result, for applications in which background noise levels are widely variable, high recognition accuracy may be difficult if not impossible to obtain.

Though only a few of the influencing factors of recognition accuracy have been mentioned, the discussion should be sufficient to illustrate how sensitive recognition accuracy can be. Studies identifying the effects of many of the factors influencing accuracy have made up a large portion of the speech recognition literature. As the potential applications increase, even more research is likely to be directed toward this purpose.

Accuracy and Task Performance

It is not difficult to understand why researchers have paid so much attention to identifying factors which affect recognition accuracy. Naturally, the accuracy of a recognizer has the potential to have a major influence on the performance of the speech recognition system. Of course, the effects of a recognition error on the task performance are entirely dependent on the application. Perhaps the most cited example of a need for near perfect accuracy levels is in the cockpit of a combat aircraft. Here, verification of all inputs is impractical, however a misrecognition could have catastrophic results and therefore initial correct recognition is necessary. To a lesser degree, Zarembo (1986) describes the need for high accuracy levels of the speech recognition system considered for use on the floor of the New York Stock Exchange. In this application, a misrecognition if not detected and corrected, could be financially devastating.

In many applications, however perfect recognition is not critical. Either errors can be corrected easily or the effects of an uncorrected error are not especially harmful to the system. The literature contains numerous descriptions of applications of voice entry systems to industrial workstations. In such applications, the operator can usually be supplied with either visual or auditory feedback, allowing him to recognize and correct any recognition errors. Similarly, in office data entry tasks errors can be corrected easily. In such applications, a small number of uncorrected errors are not particularly harmful either. In fact, the errors resulting from even a simple recognition system may be fewer in number than those resulting when more conventional input devices are used. For instance, in a simulated air traffic control task, speech input resulted in nearly 65% fewer errors than the keyboard method currently being used (Connolly, 1979).

Though recognition errors are easily corrected in many applications, task completion time will no doubt be affected due to the time required for error correction. As pointed out by Knight and Peckman (1984), some recognition errors are likely to take more time to correct than others. Substitution errors (misrecognitions) may require the longest time to correct since first the incorrect word must be deleted and next the correct utterance must be repeated. When the system fails to respond to an attempted input (either a deletion error or a rejection error) the user must simply repeat the utterance, while when an insertion error occurs, the user merely deletes the unintended input. Also, some errors may be more difficult to correct than others. Misrecognition errors, depending on the task, may result in some unintended action being performed. This may be difficult to correct. They also may be difficult to detect, since as far as the system is concerned, it has recognized a word, whereas nonrecognition errors result when the system cannot recognize an intended input. In the case of nonrecognition errors, the system can often respond with a "beep" alerting the operator that an error has been made.

Though task performance (time to complete the task, number of uncorrected errors, etc.) is generally thought to degrade as recognition accuracy declines, the exact nature of this relationship has not been investigated. Many researchers have questioned whether small increases in accuracy have enough effect on system performance to warrant the cost and trade-offs in obtaining them (Harris, 1982; Pisoni, 1986, Simpson et. al. 1985). This question is critical.because of the difficulty associated with obtaining higher and higher accuracy levels. Depending upon the effect that a lower recognition accuracy will have on the task performance, the user may choose to forgo higher accuracy in favor of some other features. For example, Craft (1982) discusses the United States Postal Service's use of speech recognition in sorting mail. He suggests that slower speaking rates would reduce the number of recognition errors. However, the cost of doing so would be slower mail delivery. In this case, system performance was improved by trading recognition accuracy for increased speaking rates. As another example, Poock et. al. (1982) found that a speaker-dependent recognition system trained by several individuals could recognize the speech of a user whose speech patterns had not previously been entrained to the recognizer (a speaker-independent condition) with a corresponding decrease in recognition accuracy of 4%. If the effects of an approximately 4% decrease in accuracy on the performance of a given task were known, the user group may find it more beneficial not to spend the time necessary to train the recognizer with each user's speech patterns and accept whatever degradation of system performance this decrease in accuracy may cause. If the effects of accuracy levels were known on task performance, similar trade-offs could be made for any number of features (e.g. high noise conditions, continuous speech, fewer training passes, etc.).

Despite the difficulty in obtaining high accuracy levels and the potential effects accuracy levels may have on task performance, few studies have attempted to look at this relationship between accuracy and task performance. Williges and Dryden (1987), in comparing two different vocabularies found that the use of one resulted in nearly sixteen times more misrecognition errors than the other. The vocabulary resulting in more errors also required subjects to spend nearly three times as long to complete the task. This study clearly indicates that accuracy level does have a tremendous impact on task completion time. The present study has attempted to better define that relationship. Subjects performed a data entry task under three predetermined levels of recognition accuracy. Measures of task performance were taken to determine what effect different levels of accuracy have on this type of task.

Accuracy and Acceptability

Though many authors have recognized the fact that the accuracy level of a recognizer plays a major role in satisfaction of the user (e.g. Nusbaum and Pisoni, 1986a; Lea, 1982), clear evidence does not exist as to how user acceptance changes as a result of various accuracy levels. Even when lower accuracy levels do not significantly affect task performance, the user may find the errors frustrating. On the other hand, small differences in accuracy level may not even be detectable to the user.

Based on observation rather than empirical evidence, Nye (1982) suggests that accuracy levels less than 90% will result in rejection by the user regardless of the application. Low levels, he states, in the 90-95% range would possibly be accepted for use with toys and games, but for use in industrial and office tasks, higher levels would be necessary. In these cases, the operator is required to perform a certain task and uses the voice recognizer as a means of assisting him. If the recognizer fails to perform well, the user may quickly lose confidence in the system and become discouraged.

Again, these guidelines are based on observation, not data collected under controlled conditions. Despite the importance of user acceptance, only one study has had as its major focus discovering how user acceptance changes as a result of accuracy level. Poock and Roland (1982) attempted to determine at what accuracy level user frustration was great enough to result in the user considering the system unacceptable. They varied the accuracy level at which the recognizer performed and had subjects read a list of task-related words. Afterwards, each subject completed a questionnaire as to the acceptability of voice input. The results of the study were disappointing, yielding only a small negative correlation between error rate and acceptability rating even though error rates involved ranged from 0% to nearly 20%. The investigators acknowledge several shortcomings in their study which may have caused what they consider to be misleading results. Namely, the task required the subjects simply to read a list of words. Although they were stopped each time an error occurred, they were not required to correct the error. By simply reading a list of words, the users did not experience the frustration that would have occurred had they been performing a realistic task with a specified goal and were interrupted by recognition errors. In addition, each subject experienced only one level of recognition accuracy and was asked to rate the acceptability of voice input. Since most of the subjects had had no previous experience with voice recognition, even those who received high error rates were probably so impressed by the fact that a machine could understand speech that they rated it favorably. Once the novelty of the device wore off, recognition errors would likely have become more frustrating. Although the results were disappointing, this study was instrumental in guiding future research efforts.

It is difficult even to extract from other investigations and case studies an idea of how users respond to speech recognizers operating at various accuracy levels. Few studies dealing with speech recognition systems report both the accuracy level at which the system operated and users' feelings about the system. As discussed previously, many studies have had the subject use a speech recognizer under various conditions (increasing noise levels for example) to determine if recognizer accuracy was affected. These studies subject the user to various recognizer accuracy levels but since the primary goal is not to determine the users perceptions, no user comments or ratings were reported. Other studies have been concerned with comparing different currently marketed recognizers (Baker, 1982; Doddington and Schalk, 1981; Nusbaum and Pisoni, 1986). Again, although a subject may experience a different accuracy level with each recognizer, no subjective measures were taken to determine how the subject felt about these differences, or even if the differences were noticeable to the user. Still other studies have compared voice input with other input methods (keyboard, mouse, etc.) and although the authors usually report the error rates associated with the voice recognition system, they have generally concentrated on objective performance measures rather than subjective ratings (Sweeney and Bitar, 1986; Taggart and Wolfe, 1981; Welch, 1977).

On the other hand, several studies have reported users' subjective ratings and comments about voice input but failed to report the average accuracy level at which the speech recognizer was operating. For instance, Biermann, Rodman, Rubin, and Heidlage (1985) report subjects expressing a favorable attitude toward a voice interactive natural language system. However, instead of recognizer accuracy, the investigators reported the percentage of successful transactions where each transaction consisted of a complete sentence. An unsuccessful transaction could be the result of one or more recognition errors within the sentence, the subject using an invalid series of inputs, or the subject changing his mind midway through the sentence and cancelling the transaction. Though for their purposes, percentage of successful transactions was clearly the important measure, it does not allow any inferences to be made about user satisfaction and accuracy levels.

In a similar manner, Poock (1980) reports very good user acceptance of a voice input system for operating a distributed computer network. In his study, however, the important measure was command input errors, whether they be due to a voice recognition error or to the subject speaking the wrong command. In a study by Morrison, Green, Shaw, and Payne (1984), subjective measures indicate that the users did not favor voice input for a text editing task, but again, the investigators did not report the accuracy level at which the system was operating.

Granted, in each case subject acceptance or rejection of the voice input system was likely due in part to a number of factors, only one of which was the accuracy of the machine. Perhaps though, some general patterns could be formed regarding acceptable accuracy levels had enough investigators reported both recognizer accuracy level and subjective ratings.

The present study has attempted to overcome the procedural limitations of the Poock et. al. study (1982) and gain some knowledge of the effects of accuracy on user satisfaction. In particular, subjects performed a realistic task with specified goals and were asked to correct all recognition errors. In addition, each subject experienced three different levels of accuracy so that they were not rating merely the acceptability of voice input, but the acceptability of voice input across different levels of recognizer accuracy.

Vocabulary Size

The second characteristic of the speech recognizer of concern in this study is the size of the vocabulary. This of course refers to the total number of different words the speech recognizer is able to identify.

Difficulties Associated with Large Vocabularies

As described earlier, nearly all currently available speech recognition systems depend on a pattern-matching procedure for recognition and are, therefore, limited to identifying words which have previously been trained and stored in memory. Current technology limits the size of this vocabulary for several reasons. First, as the size of the vocabulary increases, the necessary storage space increases. Though once a considerable limitation, advancing computing systems have nearly eliminated this as a problem. More importantly, as the vocabulary increases, the number of input-template comparisons which must be made before a best match can be determined also increases. As a result, greater computational power is required in order to keep the system response time from escalating (Kurzweil, 1986).

Perhaps the most limiting problem of larger vocabulary sizes is the corresponding decrease in recognizer accuracy. As has been explained, perfectly accurate recognition is not always critical, and the exact effects of lower accuracy levels on system performance are not known. However, for any applications reasonably good recognition is required, and large vocabularies often result in very low accuracy levels.

Based on the template matching procedure of recognition, as the vocabulary size increases, the chance of misrecognition errors also increases due simply to the increase in the number of comparisons which must be made. Most importantly, as the vocabulary size increases, the number of acoustically similar words within the vocabulary naturally increases.

The inherent confusability of vocabulary items has been shown to be one of the primary factors influencing recognition accuracy. Vocabularies consisting of a large number of acoustically similar items (i.e. "pen" and "ten", "bottle" and "bottom") pose a problem for speech recognizers. This is clearly illustrated in a study which used both the orthographic (written) alphabet and a phonetic alphabet (alpha, bravo, charlie, ...) (Schurick, 1986; Williges and Dryden, 1987). The orthographic alphabet, of course, consists of a large number of phonetically and acoustically similar words (e.g., b, c, d, e, g, p, t, v, z). In a data entry task, use of the phonetic alphabet resulted in approximately 15% greater recognition than the same task performed using the orthographic alphabet. Similarly, in developing empirical prediction equations for speech recognizer performance, Spine, Williges, and Maynard (1984) first developed and tested two vocabulary types, an inherently confusable vocabulary and an acoustically distinctive vocabulary. As expected, performance differed drastically on the two, leading the researchers to develop separate models for each vocabulary type. In a series of studies (Levison, Rabiner, Rosenberg, and Wilpon, 1979; Rabiner and Wilpon, 1979) investigators showed that despite increasing the size of a vocabulary, better recognition can be obtained by reducing the similarity of the vocabulary items. Accuracy on a vocabulary of 39 acoustically similar words was only about 80% while on a larger vocabulary (54 words) consisting of more distinct words, accuracies as high as 98% were obtained. Inevitably, as vocabularies exceed one to two hundred words, attempting to include only acoustically distinctive words becomes a futile effort.

Because of the difficulties associated with large vocabulary speech recognition systems, most of the recognizers available today are limited to vocabularies of 100 to 200 words. A few higher-priced systems are capable of identifying as many as 1000 words. However, their high cost is prohibitive for many applications, particularly as computer entry devices for the disabled. These limited vocabulary recognizers have proven useful for a number of applications. Other tasks, which may possibly be aided by voice input, require more extensive vocabularies than can be accommodated by 100 to 200 words. For example, one of the yet unrealized goals of many researchers is to develop an automatic dictation machine. A user simply composes a letter or report aloud and immediately receives a printed copy. For such applications, clearly a much larger vocabulary is required.

Expanding Vocabularies through Syntax Control

One method of extending the capabilities of smaller vocabulary recognizers is through the use of syntax structure (Fink, 1980; Nye, 1982). When the task involves entering a logical sequence of inputs, the choice of acceptable words at each point in the task is limited. By partitioning the total vocabulary into a set of smaller vocabularies corresponding to the acceptable input choices at each point in the task, a large vocabulary may be accommodated by a recognizer with a modest vocabulary size. The total vocabulary may consist of several hundred words which are stored in memory away from the actual recognizer. When a subset of this vocabulary is required, it is downloaded into recognizer memory and then uploaded again when no longer needed.

To illustrate the use of syntax control, suppose a mail order catalog company uses voice recognition to accept orders. After entering the item name, the buyer may be prompted for the color he wishes. At this point, the only acceptable inputs would be the names of colors. This subset of the vocabulary becomes active, the user inputs the desired color name, then the subset of color names is returned to storage. The next item to enter may be the quantity. At this point, the subset of digits (0-9) becomes the active vocabulary and so on. This is a very basic example of syntax control. More complicated syntax structures may be developed which, for example, use the rules of grammar to restrict word choices at each point in an English sentence.

Syntax control, however is not without its limitations. Namely, the user is required to remember where he is in what can become a complex hierarchical arrangement (Nye, 1982). If he forgets which branch of the syntax tree he is in, the system may reject everything he says because he is using words from an inactive vocabulary subset. If the software provides for sufficient prompting and feedback, this problem may be minimal; otherwise the user must pay careful attention to which portion of the structure he enters, which increases the task workload on the user. In applications which use grammatical syntax control, other problems arise. For example, when composing a letter the user may first wish to dictate a rough draft in order to get his ideas in print quickly. It is doubtful in such a case that the user will always use grammatically correct English sentences. Being forced to comply with the required syntax structure may cause substantial interference with the thought processes (Meisel, 1986).

In some situations, syntax control simply cannot be used as a means of increasing vocabulary size. Clearly, to use syntax control, the task must be one in which the inputs are structured in some fashion. As mentioned before, dictating an unformatted rough draft would not lend itself to the use of syntax control. In addition, some sub-vocabularies are too large for a recognizer to handle, yet cannot be divided any smaller. And there are situations where the possible input items are not predictable such as all of the possible customer names in the mail order catalog example used earlier.

Expanding Vocabularies thru Character-level Entry

A second method of increasing the number of possible inputs does not suffer from these limitations. This method uses the speech recognizer as a voice activated keyboard. The user enters text character-by-character rather than word-by-word. Most frequently used words can be entered at the word level, while any remaining words can be entered by spelling them. Even as technology allows recognizer vocabularies to increase (see for example Kurzweil, 1986) the possibility always exists that a new word not previously trained will need to be entered. Also, because of the time required to train larger and larger vocabularies (Nusbaum and Pisoni, 1986), it may be more efficient to enter a large number of infrequently used words by spelling them rather than training all of them in advance even when the recognizer has the capacity to do so.

Effects on Task Performance

The efficiency of this method of speech entry has been investigated only to a limited degree. Spelling words is of course slower than entering complete words. One study found that in a task consisting of text, digits, and control words, word-level entry resulted in an average entry rate of 32.5 words per minute while character level entry was slower at an average rate of 22.5 wpm (Schurick, 1986; Williges and Dryden, 1987).

Naturally, the smaller the recognizer's storage capacity, the more words must be entered by "spellmode". Therefore, time to complete a task will no doubt increase with smaller recognizer vocabulary sizes. The nature of this relationship was investigated in part by Gould, Conti, and Hovanyecz (1983). The purpose of the study was to determine the feasibility of a limited capability automatic dictation machine, or as they called it, a listening typewriter. In doing so, they simulated along with isolated and connected speech modes, various vocabulary sizes. In the first experiment users composed and edited letters with the simulated voice recognizer which had either a 1000 word vocabulary or an unlimited vocabulary. The 1000 word vocabulary was composed of the 1000 most frequently used English words. An analysis afterwards indicated that roughly 75% of all of the words used in the letter writing task were available in the 1000 word vocabulary, while the remaining 25% were entered by spelling them. The unlimited vocabulary, of course, required no spelling. Task completion time was found to be slower when users were required to enter a portion of their inputs by spellmode. To compose and edit a letter using the 1000 word vocabulary required almost 30% more time than when the unlimited vocabulary was used (averaged for both isolated and continuous speech modes). Time to comnose a letter using isolated speech and the 1000 vocabulary may in fact take even longer than this study indicates. In composing letters using isolated word entry, subjects were required to leave a distinct pause between words just as would be necessary with an actual isolated-word speech recognition system. However, when the subjects entered spellmode, they were no longer required to speak in isolated utterances. It is not clear why the researchers chose to simulate the system in this manner, because most isolated speech recognition systems would require isolated speech at all times. So in fact, there is likely an even greater difference in task completion time for the two vocabulary sizes than this study indicates.

In the second experiment, in addition to the 1000 word and unlimited vocabulary, a 5000 word vocabulary was also investigated. This vocabulary was found to include approximately 90% of all the words used by the subjects in the letter writing tasks, therefore only 10% of the inputs were entered character-by-character. Results indicated a smaller difference between task completion times for the 5000 word vocabulary condition and the unlimited vocabulary condition. Under the 5000 word vocabulary condition, this group of subjects required approximately 17% longer to complete the task than when using the unlimited vocabulary, and nearly 40% longer when using the 1000 word vocabulary. These results are for the isolated speech condition only because in order to reduce the complexity of the design the researchers did not evaluate the 5000 word vocabulary in conjunction with the continuous speech mode. As in the first study, when using the isolated speech condition, subjects were not required to speak in isolated utterances during spellmode. Therefore, the task completion times for the limited vocabulary conditions would likely be even greater in comparison to the unlimited vocabulary conditions.

In another study, Williges and Dryden (1987) varied the amount of character-level entry required for a data entry task. In what they called the efficient dialogue, approximately 10% of the input was entered by spellmode while on the same task using what they term the inefficient dialogue, somewhere between 20 and 30% of the entries were spelled. Using the inefficient dialogue, subjects required nearly three times as long to complete the task as with the efficient dialogue. These results however, are confounded with another factor. As mentioned earlier, one of the primary factors of interest in this investigation was the difference in recognition accuracy and entry speed between the phonetic and the orthographic alphabets. Spelling under the efficient dialogue used a phonetic alphabet while spelling under the inefficient dialogue used the orthographic alphabet. Naturally, because of the increased recognition errors when using the orthographic alphabet (nearly 15% more errors), much of the difference in task completion time was due to the increased amount of error correction necessary rather than the increased amount of character-level entry required. Unfortunately, the results do not allow inferences to be made regarding what portion of the time difference was due to each factor.

In the present study, subjects performed a data entry task using speech input. In each trial, they used one of three vocabularies differing in the percentage of words which could be entered by word-level entry. Time to complete the task and number of errors were collected and analyzed in order to determine what, if any, effects vocabulary size has on these measures.

Effects on User Satisfaction

The Gould et. al. (1983) study also measured user's attitudes regarding the different recognizers they used. In the first experiment, after each condition the subject would rate the recognizer on a seven point scale, anchored on either end by "significantly worse than writing (1)" and "significantly better than writing (7)". The conditions which used a 1000 word vocabulary averaged a 5.0 preference rating while the unlimited vocabulary conditions averaged a 5.75 rating (averaged across both isolated and continuous speech modes). Subjects' comments indicated that having to stop and spell words interrupted their train of thought but if they waited until the end of the task to correct the unidentified words, they often forgot what they had intended to say. Had the subjects been required to speak in isolated utterances while spelling words during the isolated speech conditions, they may have found the constrained vocabulary even more disrupting to their task and may possibly have rated it lower.

In the second experiment, the subjects were all experienced dictators, and instead of comparing the voice input to writing, they compared it to their favorite method of composing letters whether that be writing, dictating to a machine, or dictating to a secretary. Again, a seven-point scale was used. On average, subjects rated the 1000 word and the 5000 word recognizers less favorable than their preferred method (2.25 and 3.5 respectively), while the unlimited vocabulary conditions received a slightly more favorable rating (5.0) (averaged across both speech modes). Remember however, that the 5000 word vocabulary was not used in conjunction with the continuous speech mode which generally was rated more favorable than isolated speech. Therefore the average rating for the 5000 word vocabulary condition under both speech modes would likely be higher and may have even surpassed the neutral rating of "the same as my favorite method (4)".

In the present study, each subject used the speech recognition device for a data entry task. Three levels of available vocabulary size were investigated. After each treatment condition, the subject rated the device on a number of parameters. These parameters have been analyzed in order to discover the effects of different vocabulary sizes on the acceptability of voice input.

Age

Speech recognition systems may well find their way into a wide variety of applications including not only those in which a small group of workers will share the system, but also those which will be available to the general public. An example of such a system would be an automated teller machine which recognizes the speech of its users. As such, it is necessary to determine how different age groups, including those not usually found within the same working communities, respond to speech recognition systems.

It is well known that age is a significant factor influencing performance in many tasks. However, Ogozalek and Praag (1986) investigated the effects of age on performance of a letter composition task using speech input and found no significant differences between younger and older users. The letter composition task was similar to that used by Gould et. al. (1984), requiring subjects to enter and edit letters. Older subjects performed as quickly, accurately, and effectively as the younger subjects. It should be noted, however, that the older subjects used in this study were at the time enrolled in a computer course, and in fact may not be representative of the general population of older persons. In addition, subjects in the study were not required to make a large number of error corrections because the simulated "listening typewriter" had an unlimited vocabulary and did not misrecognize words. A study by Rosson and Mellon (1985) indicates that older adults may in fact have more difficulty in correcting errors in such a task than younger ones, which ultimately means task performance would be poorer for older persons when error correction is involved. Their study involved a data retrieval task using a hierarchical data structure as well as synthetic speech. Younger subjects were able to use the system much more quickly and more accurately, primarily because they had less difficulty correcting errors. The task in this study was not very similar to a speech input task, however, the procedure for correcting errors was. Therefore, it is not unreasonable to propose that higher misrecognition rates may be more detrimental to system performance for older than younger people.

The age of a person has been shown to affect significantly his/her perceptions of high technology equipment as well (Clark, 1986). Different age groups of users may not only have different views concerning speech input in general, but may also have different expectations as to the capabilities a system must have before it is considered acceptable. The Ogozalek and Van Pragg (1986) study measured users' attitudes toward the "listening typewriter" and found that older subjects were much more enthusiastic about using such a system. As mentioned before, this study did not simulate different vocabulary sizes or accuracy levels and therefore subjects experienced a near-perfect speech recognition device. The Poock and Roland (1982) study investigated the relationship between age of the subject and the acceptability of different accuracy levels. No relationship was found, but because of the problems associated with the study (as discussed before), these results are not conclusive. For applications which attract a wide range of users, the system should be designed to accommodate all possible users. For this reason, it is necessary to know how different age groups of people respond to voice recognition.

The present study involved three groups of subjects ranging in age from 20 to 55 years. Objective and subjective measures have been analyzed with respect to age in order to determine whether these measures are dependent on the age of the user.

Method

Subjects

Three groups of participants with six subjects in each group (for a total of 18 subjects) were sought from the university community to participate in this study and were compensated for their time. The youngest group consisted of subjects between the ages of 20 and 25 years. The intermediate age group was comprised of subjects between the ages of 35 and 40 years, while the older group consisted of subjects between the ages of 50 and 55 years. These particular age groups were chosen because they represent a wide range of users and because subjects within these ages are readily available within a university community. Each group was equally divided across gender. All participants were required to be native speakers of English, to have no detectable speech impediments, and to have had no previous experience with voice recognition. The questionnaire used to screen potential subjects is presented in Appendix A.

Experimental Apparatus

Because of the need to manipulate experimentally the operating accuracy level and vocabulary size, a simulation of a speech recognizer rather than an actual recognizer was employed (to be discussed in detail in the Experimental Procedure section). This simulation used two Digital Equipment Corporation VT220 terminals connected to a VAX 11/750 mainframe system operating under VMS, and a specially developed PASCAL program. A General Electric video cassette recorder was used to record both audio and video signals.

Experimental Design

The experiınental design consisted of a 3x3x3 mixed-factor factorial design. The design matrix appears in Figure 1. This design involves three independent variables: recognizer accuracy, available vocabulary, and user age.

Recognizer accuracy is a fixed-effects, within-subject variable. As shown in Figure 1, three levels were investigated: 91%, 95%, and 99%. As discussed previously, recognizer accuracy refers to the number of spoken vocabulary items which are correctly identified by the recognition system. In the 91% recognition condition for example, 9 of every 100 spoken vocabulary words were not identified by the system.

Available vocabulary is a fixed-effects, within-subject variable. For this study, the available vocabulary refers to the percentage of words needed for the task which

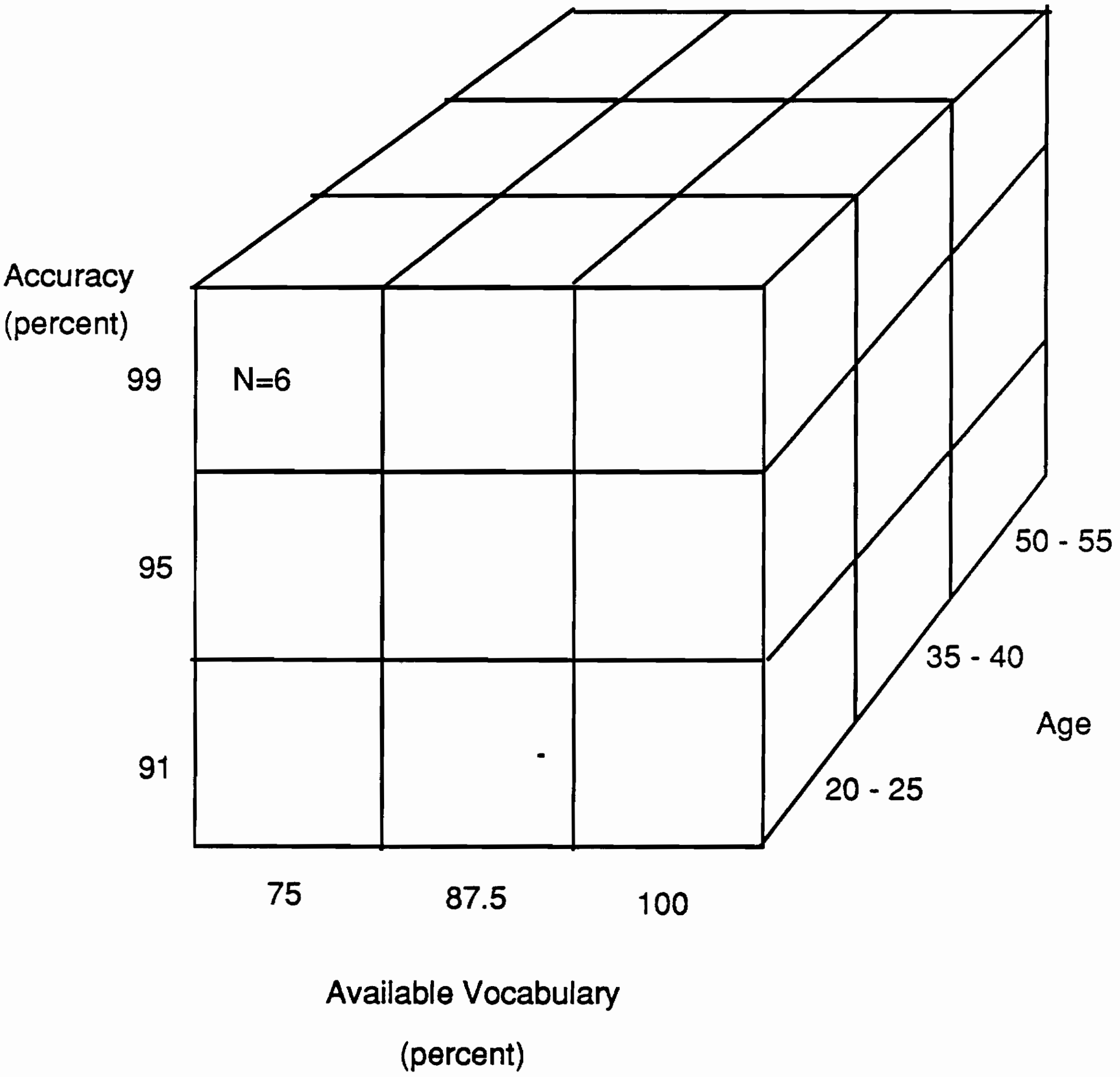


Figure 1. Experimental Design: A 3x3x3 within subject factorial design

appear in the recognizer's vocabulary. For instance, if all the words which are needed in the task have been previously trained and are available for use, the available vocabulary is referred to as 100%. If however, the recognizer is not able to accommodate as many words and for example, 20% of the needed words must be entered by spelling them, then the available vocabulary is referred to as 80%. This measure, as opposed to simply the number of words in the recognizer's vocabulary, was chosen because it describes the recognizer vocabulary with respect to the task requirements. After all, its the difference between the number of words required for a task and the number of words the recognizer is able to discriminate that indicates the amount of character-level entry required. This measure, then, is generalizable to tasks requiring any vocabulary size. An 80% available vocabulary could equally describe a recognizer with a 100 word vocabulary applied to a 125 word task or a recognizer with a 500 word vocabulary used for a task requiring 625 different words. In each case, 20% of the inputs will be made using character-by-character entry. This study investigated three levels of available vocabulary: 75%, 87.5%, and 100%.

Subject's age is a fixed-effect between-subject variable with three levels: young (20-25 years), intermediate (35-40 years), and older (50-55 years). Because the third factor, age, is a between-subject variable, each subject experienced 9 treatment conditions (resulting from the three levels of recognizer accuracy and the three levels of vocabulary size). By using 18 subjects, it was possible to balance the presentation order of these 9 treatment conditions using a Latin Square arrangement. Because the number of treatment conditions is odd, two Latin Squares were used to balance the design, thereby requiring 18 subjects, six from each of the three age groups. Table 2 illustrates the presentation order of the 9 treatment conditions to each of the 18 subjects.

Table 2. Treatment condition presentation order

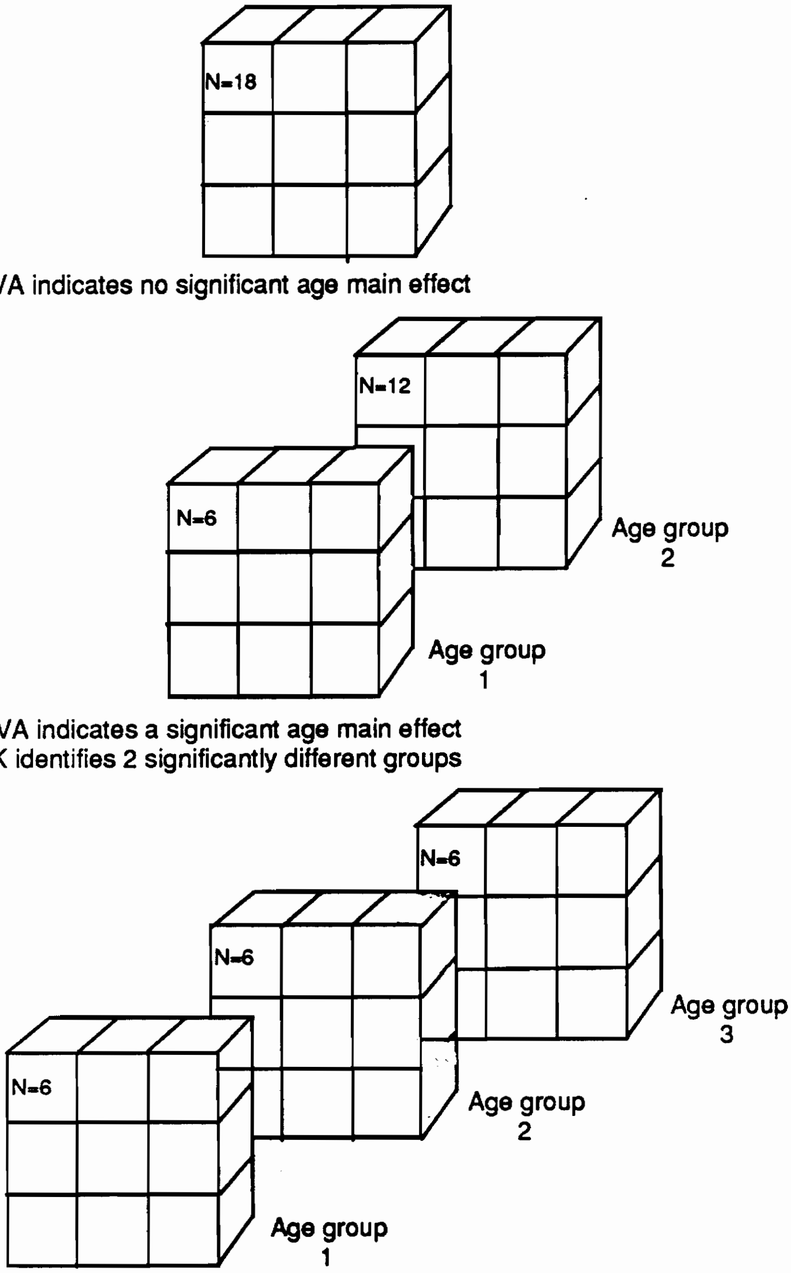
Subject																	
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	2	3	4	5	6	7	8	9	5	6	7	8	9	1	2	3	4
9	1	2	3	4	5	6	7	8	6	7	8	9	1	2	3	4	5
2	3	4	5	6	7	8	9	1	4	5	6	7	8	9	1	2	3
8	9	1	2	3	4	5	6	7	7	8	9	1	2	3	4	5	6
3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2
7	8	9	1	2	3	4	5	6	8	9	1	2	3	4	5	6	7
4	5	6	7	8	9	1	2	3	2	3	4	5	6	7	8	9	1
6	7	8	9	1	2	3	4	5	9	1	2	3	4	5	6	7	8
5	6	7	8	9	1	2	3	4	1	2	3	4	5	6	7	8	9

where treatment	1	corresponds to	(voc 75, acc 91)
	2		(voc 75, acc 95)
	3		(voc 75, acc 99)
	4		(voc 87.5, acc 91)
	5		(voc 87.5, acc 95)
	6		(voc 87.5, acc 99)
	7		(voc 100, acc 91)
	8		(voc 100, acc 95)
	9		(voc 100, acc 99)

A 3 x 3 complete factorial design can be viewed as a special case of a second-order response surface methodology central-composite design (Cochran and Cox, 1957). It is considered a special case because it meets the conditions of a central composite design, yet is still a complete factorial thus also allowing for an analysis of variance. The design used in this study was chosen so that it could be considered as one or more 3x3 factorial designs. If initial analysis were to indicate no significant age main effect, the design would be collapsed across age to form a single 3x3 within subject design (accuracy x available vocabulary) with 18 subjects assigned to each treatment condition. If initial analysis found the age main effect significant, then the design could be considered a 3x3 within-subject design replicated either two or three times, depending upon whether post-hoc tests indicate two or three significantly different age groups. In the case that two significantly differently age groups are identified, the result would be two 3x3 designs with 6 subjects assigned to one and 12 subjects assigned to the other. For three significantly different age groups, the design would then be viewed as a 3x3 design replicated three times, with 6 subjects assigned to each. Figure 2 illustrates this idea.

Dependent Variables

In order to understand the influence of recognizer accuracy and available vocabulary on the usefulness of voice input, two types of data were collected from each subject. Objective measures were collected and used to determine the effects of recognizer accuracy and available vocabulary on task performance. Subjective measures were



a) ANOVA indicates no significant age main effect

b) ANOVA indicates a significant age main effect
S-N-K identifies 2 significantly different groups

c) ANOVA indicates a significant age main effect
S-N-K identifies 3 significantly different groups

Figure 2. Experimental design viewed as (a) a single 3x3 design (b) a 3x3 design replicated twice or (c) a 3x3 design replicated three times

collected and analyzed to determine the effects of recognizer accuracy and available vocabulary on the subjects' judgement of acceptability of voice input.

Task completion time was the primary objective measure of interest. This time was measured from when the subject spoke the first input to when the recognizer displayed the last spoken word of the task. This includes time to enter data, detect recognition errors, delete those incorrect entries and repeat the correct word. Because the possibility existed that a subject would not detect a recognition error or an invalid vocabulary item, the number of uncorrected errors remaining at the end of a session was also recorded for each session. (The exact types of errors which may occur are discussed in the Experimental Procedure Section.) To form a more meaningful measure, the percentage of incorrect words remaining at the end off each trial was calculated. This measure is simply the number of incorrect words divided by the total number of words, and is referred to as Percent Final Error or simply as the final error rate.

In addition to performance measures, subjective satisfaction with the voice recognition system was assessed following each treatment condition. The instrument used for subject evaluation of the voice input system was a set of 13 bipolar rating scales. These 13 bipolar scales consisted of an overall acceptability scale and 12 individual adjective scales. These adjectives and the anchors for the accompanying seven-point scale are presented in Table 3. This multidimensional approach, as opposed to using a single acceptable/unacceptable scale, is based on work by Osgood, Suci, and Tannenbaum (1957) who suggest that results obtained in such a manner are more representative and reliable than when only a single measure is used. The specific adjectives used were chosen based on a scale developed by Coleman (1985)

Table 3. Subjective rating scale

Acceptable						Unacceptable
extremely	quite	slightly	neutral	slightly	quite	extremely

Fast	Slow
Accurate	Inaccurate
Natural	Unnatural
Complete	Incomplete
Comfortable	Uncomfortable
Consistent	Inconsistent
Pleasing	Irritating
Dependable	Undependable
Friendly	Unfriendly
Facilitating	Distracting
Simple	Complicated
Useful	Useless

to evaluate the interface of a text editor. In order to be more applicable for evaluating a speech input system, however, it was necessary to modify the scale. Although Coleman clearly describes the procedure used in constructing his scale, the procedure could not be implemented here because it relies heavily on querying experienced users of systems similar to that which is to be evaluated. Because speech input systems are not in wide-spread use, experienced users are not common. Therefore, the bipolar adjective scale used in this study is simply a modification of the existing scale rather than a complete reconstruction.

Experimental Procedure

Before a detailed account of the experimental protocol is presented, a discussion of the method for manipulating accuracy level and vocabulary size follows.

Accuracy and Available Vocabulary Control

As described before, this study considered three levels of accuracy as well as three levels of available vocabulary. Because of the enormous number of factors which affect recognition accuracy, it is not possible to predict, much less manipulate the accuracy at which a recognizer will operate. In addition, this study considered a very high accuracy rate (99%) and a large vocabulary. Both of these features are difficult if not impossible to obtain from currently available speech recognition systems. For these reasons, a simulation of a speech recognizer rather than an actual speech respeech recognition system was developed and used. This method of studying recognition systems has become popular within the past few years because it gives the researcher the capability of experimentally varying different recognizer features as well as evaluating new systems before they are fully developed (e.g. Cole, 1986; Gould et. al., 1983; Holmgren, 1983; Kinkead, 1986; Poock and Roland, 1982; and Zoltan-Ford, 1984).

The simulation used in this study is based in part on that used by Gould et. al. (1983). From the subjects' perspective, the simulation appeared quite realistic. He/she spoke into a microphone and the word spoken appeared on a CRT display terminal positioned in front of him. In actuality however, the microphone led not to a voice recognition machine, but to a skilled typist seated in an adjacent room. As the subject spoke, the typist typed exactly what was spoken. This information appeared not only on the typist's computer terminal, but also on the subject's terminal which was yoked to the typist's through the system mainframe. A video cassette recorder was connected to the typist's terminal screen and to the microphone in order to record each experimental condition.

Because accuracy levels of less than 100% and available vocabularies of less than 100% were being investigated, not everything the subject said was "recognized". Therefore, a method was necessary to determine which spoken words were indeed acceptable vocabulary items and which inputs would be "misrecognized". Clearly, if quick system response were to be obtained, this task could not be the responsibility of the typist. Therefore, for this purpose, a computer program was developed to accept the typist's input and determine the feedback sent to the subject. As the typist enters a spoken word, the program first compares the input word with the database of acceptable vocabulary items. The contents of this list of course changed during the study depending upon the available vocabulary level being simulated. For the 100% vocabulary condition, the list contained all the words necessary to complete the task. Even in this condition it was necessary to compare the spoken word with a vocabulary list to avoid accepting an unintended input which should clearly not be contained in the task vocabulary (e.g. the subject speaking to himself or to the experimenter during the task). The 87.5% available vocabulary comprised fewer words, and the 75% vocabulary even fewer. When the subject attempted to input a word not contained in the acceptable vocabulary, the recognizer returned instead the symbol XXXX indicating to the subject that the spoken word was not recognized.

To simulate the various accuracy levels, the program randomly selected a number of words to be either misrecognized or not recognized. The number of words selected depended on the accuracy level being simulated. For example, during the 91% accuracy condition, 9 of every 100 spoken inputs were selected by the program as "recognition errors". Of the recognition errors, approximately one third were substitution errors and two thirds were nonrecognition errors. To simulate a misrecognition, the program returned a predetermined runner-up word to the subjects' screen rather than the spoken word. For each vocabulary item, a runner-up word which is acoustically similar to the given word was defined. To simulate a nonrecognition, the program returned the symbol XXXX. No distinction was made between a nonrecognition and an invalid vocabulary item because in actuality, a true voice recognition system would not be able to distinguish the two.

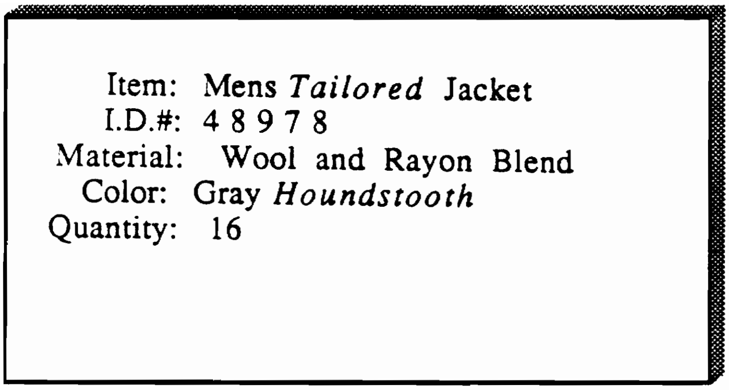
Data Entry Task

The task itself involved using speech input to enter information similar to that needed in an apparel store inventory control system. The information to be entered included the item name, the item stock number (5 digits), the quantity, and other descriptive information depending on the product (i.e. color, fabric, etc.). As mentioned, three levels of available vocabulary were investigated. In the 100% available vocabulary, all of the inputs were made by word-level entry. For the smaller available vocabulary levels, a portion of the words were chosen as not belonging to the recognizer's vocabulary and therefore required character-level entry. The words chosen were those which would not be expected to appear frequently in such an application. For example, for the 75% vocabulary condition, 25% of all the words to be entered were not included in the allowable vocabulary. These words were those which probably would not appear very often if the whole possible range of products were considered. The digits (0-9), popular colors, and general categories of merchandise (e.g. blouse, shoes, coats, etc.) were included while less frequently used words (e.g. turquoise, spandex, argyle, etc.) required character level entry.

In such an application, it is reasonable to assume that an experienced operator has a fairly good idea of which words are in the vocabulary and which must be entered character-by-character and would therefore avoid first speaking words which must be spelled. However, the subjects in this study were naturally not familiar enough with the task to know the vocabulary limitations and would therefore probably speak each word first, then spell those which were not recognized. In order to represent an actual task more realistically, the inventory list the subject received was color-coded indicating which words may be spoken (black) and which words must be spelled (red).

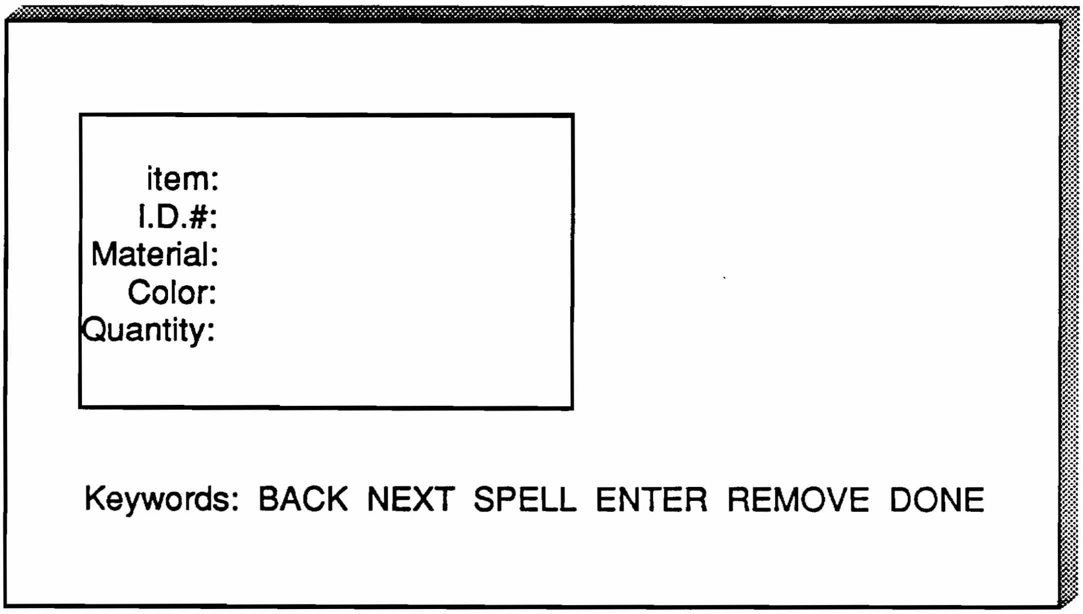
The subject received a stack of inventory cards and was told that these cards represented the shipping forms accompanying each item arriving at the department store's warehouse. A sample illustration of a card is presented in Figure 3. Each card contained the item name, stock number, quantity included, and descriptive information. As explained to the subject, all of this information was to be entered into the computer so that the company can keep a record of the inventory they receive. They were to use voice input to enter the information and were provided a CRT display for both prompting and visual feedback. An illustration of this display screen is also shown in Figure 3. Subjects first spoke the item name and then said "enter" to move to the next data field where they spoke the stock number, digit by digit. They continued until all information on a single card had been entered and then said "done" to transmit the information and proceed to the next inventory card. Each of the 9 experimental trials consisted of 15 different inventory cards. The number of words varied from card to card, but the total number of words contained on a set of 15 cards was 225 words. Therefore, the same amount of information was entered under each of the 9 treatment conditions.

While entering information, misrecognitions and nonrecognitions occurred frequently (depending upon the condition, of course). In the event of a misrecognition error, the subject would say "back" to erase the incorrect word and then repeat the item. The same procedure was used to correct a nonrecognition, saying "back" to erase the XXXX symbol, and then repeat the entry. When subjects encountered a word printed in red (meaning the word was not contained in the recognizable vocabulary), they first



Item: Mens *Tailored* Jacket
I.D.#: 4 8 9 7 8
Material: Wool and Rayon Blend
Color: Gray *Houndstooth*
Quantity: 16

Example inventory card
* italicized words represent those which appeared
in red print on the subject's card



item:
I.D.#:
Material:
Color:
Quantity:

Keywords: BACK NEXT SPELL ENTER REMOVE DONE

Illustration of subject's screen

Figure 3. Sample data card and illustration of subjects' screen

said "spell" which then allowed them to enter the word character by character. They then said "end spell" to return to word-level input. As seen in Figure 3, the list of commands were displayed on the terminal screen at all times to provide a prompt for the subject. While the subject was in word-level entry mode, the commands consisted of enter, done, back, next, and spell. When the subject entered character-level entry, the choice of acceptable commands decreased and only back, next, and end spell were active. Recognition errors were simulated not only while entering words, but also while spelling (i.e. print an "e" when a "b" was intended) and while speaking commands. Subjects corrected these errors in a similar manner. An even more detailed description of the data entry and error correction procedures can be found in Appendix C.

Protocol

Each subject participated in two experimental sessions, each lasting approximately 2 hours and fifteen minutes. The first session began by having the subject read a task description, read and sign an informed consent form, and review the subjective rating scales. He then received a practice trial in which he entered inventory information from a set of 15 data cards, while the experimenter answered any questions which may have arisen. Next, the actual experimental trials began. The first day, each subject completed 4 of the 9 treatment conditions. The subject was instructed to speak in a "connected-speech" mode and was asked to speak more rapidly or more slowly by the experimenter if necessary. This was rarely ever necessary, however, because each subject seemed to adopt a pace that would allow him to monitor the feedback and correct any errors before the next word was entered. In addition, it was stressed that subjects speak each word clearly. After each trial, the subject responded to the subjective rating scales. On the second day, each subject received the remaining five treatment conditions, was debriefed, and paid for participation.

Data Analysis and Results

Task Completion Time

The time required for each subject to complete the task under each of the 9 treatment conditions was recorded (in seconds). Again, this time was measured from when the subject spoke the first word on the first data card to when the system displayed the last spoken word of the last data card of the set.

An ANOVA was run on the resulting times using Accuracy, Available vocabulary, and Age as the main effects. These results are shown in Table 4. The main effects of both Accuracy and Vocabulary were significant ($p < 0.001$), as was their interaction ($p < 0.001$). The main effect of Age was not found to be significant, ($p > 0.05$), however the interaction between Age and Accuracy was ($p < 0.05$). A Student-Newman-Keuls test was performed on the means of the significant effects as shown in Table 5.

Table 4. Anova for Task Completion Time (Combined Age Groups)

Source	*df*	*SS*	*F*	*Prob*
Between Subjects				
Age	2	151057.64	1.01	0.3874
Subjects (S)	15	1120892.09		
Within Subjects				
Acc	2	2008349.86	132.79	0.0001
Acc x Age	4	121240.62	4.01	0.0101
Acc x S(Age)	30	226855.74		
Voc	2	7720077.64	159.06	0.0001
Voc x Age	4	116091.06	1.20	0.3329
Voc x S(Age)	30	728039.52		
Acc x Voc	4	256631.84	6.22	0.0003
Acc x Voc x Age	8	90403.12	1.10	0.3790
Acc x Voc x S(Age)	60	618871.48		
Total	161	13158510.62		

Table 5. Student-Newman-Keuls Test for Task Completion Time (Combined Age Groups)

Voc	*Mean*		*Acc*	*Mean*	
75	1139.39	A	91	1101.00	A
87.5	871.19	B	95	865.78	B
100	604.67	C	99	738.46	C

Acc x Voc	*Mean*		*Acc x Age*	*Mean*	
91 x 75	1331.39	A	91 x 3	1064.78	A
95 x 75	1157.61	B	91 x 2	1032.72	A
91x 87.5	988.83	C	91 x 1	935.50	B
99 x 75	929.17	C	95 x 3	921.44	B C
95 x 87.5	858.78	D	95 x 2	862.50	C D
99 x 87.5	765.94	E	95 x 1	813.39	D
91 x 100	712.78	E	99 x 1	749.67	E
95 x 100	580.94	F	99 x 3	736.11	E
99 x 100	520.28	F	99 x 2	729.61	E

*Means with the same letter are not significantly different from each other at $p < 0.05$

All three levels of available vocabulary were significantly different from one another as were all three levels of accuracy, with task completion time increasing with a decrease in either available vocabulary level or accuracy.

As noted before, a significant age main effect for task completion time was not found, however, the Age x Accuracy interaction was found to be significant. As such, it would likely be worthwhile to analyze the data collapsed across age as well as separated into the three age groups. Figure 4 shows the mean task completion time by accuracy and available vocabulary for the collapsed data while Figures 5, 6, and 7 show the same for each separate age group. In each case, a reduction in either available vocabulary or accuracy level produces an increase in task completion time.

As shown in Tables 6, 7, and 8, ANOVA's were then performed on each age group individually. In each case, the main effects of Accuracy and Available vocabulary were significant ($p < 0.001$) with post-hoc Student-Newman-Keuls tests indicating significant differences between all three levels of Available vocabulary and all three levels of Accuracy, as shown in Tables 9, 10, and 11. The interaction of Available vocabulary and Accuracy was not significant for the youngest age group ($p > 0.05$) but was for age group 2 ($p < 0.001$) and age group 3 ($p < 0.05$). The student-Newman-Keuls tests reveal slightly different loci of significance for the two groups as illustrated in Tables 10 and 11.

As described before, the design used in this study is suitable for response surface analysis. In this study, response surfaces were generated using the RSREG procedure in SAS (SAS, 1986). RSREG generates a quadratic surface using the method of squares. The procedure's output includes the regression equation, the coefficient of determination (R-squared), lack of fit, and the significance of the linear, quadratic, and

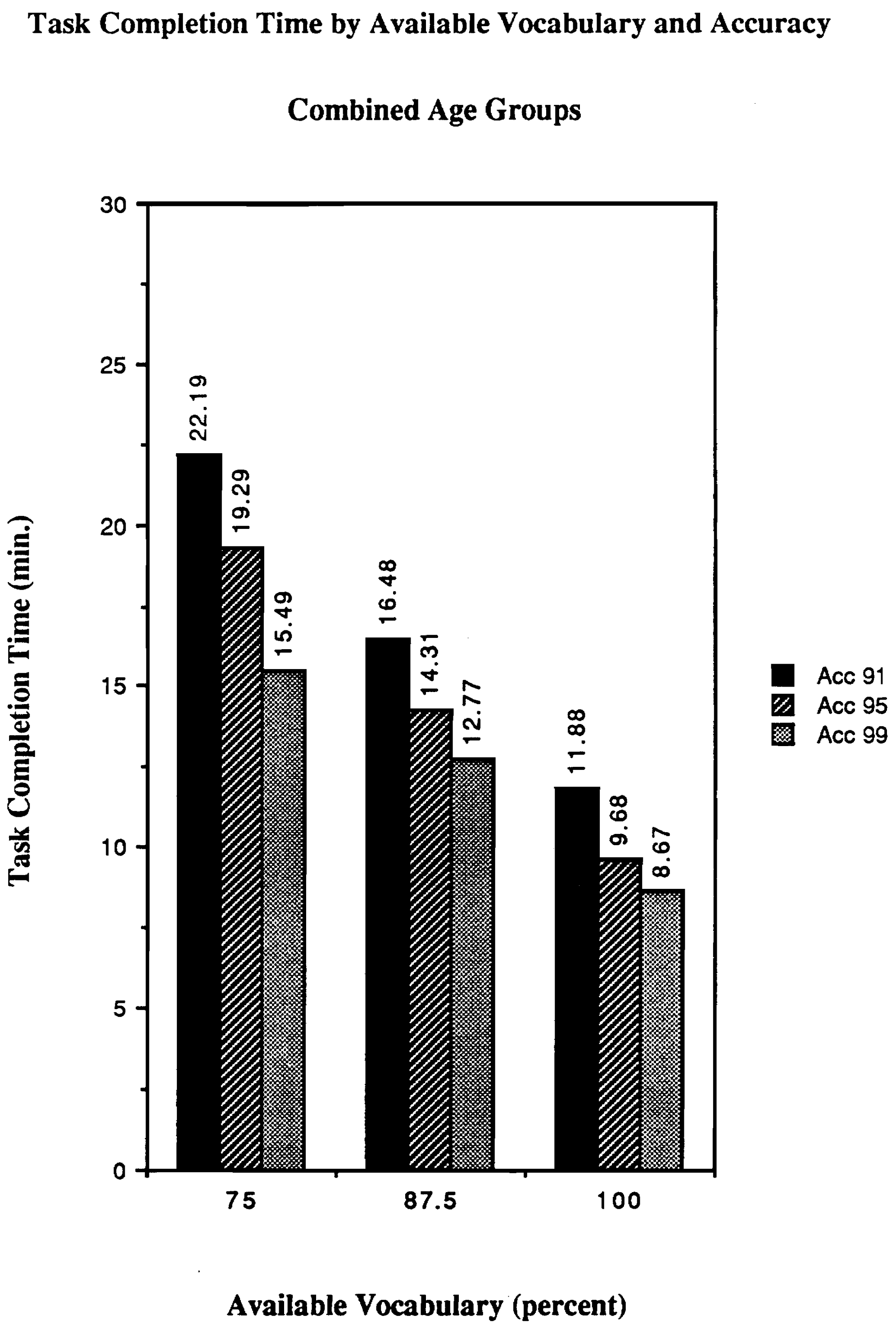


Figure 4. Task Completion Time by Available Vocabulary and Accuracy (Combined Age Groups)

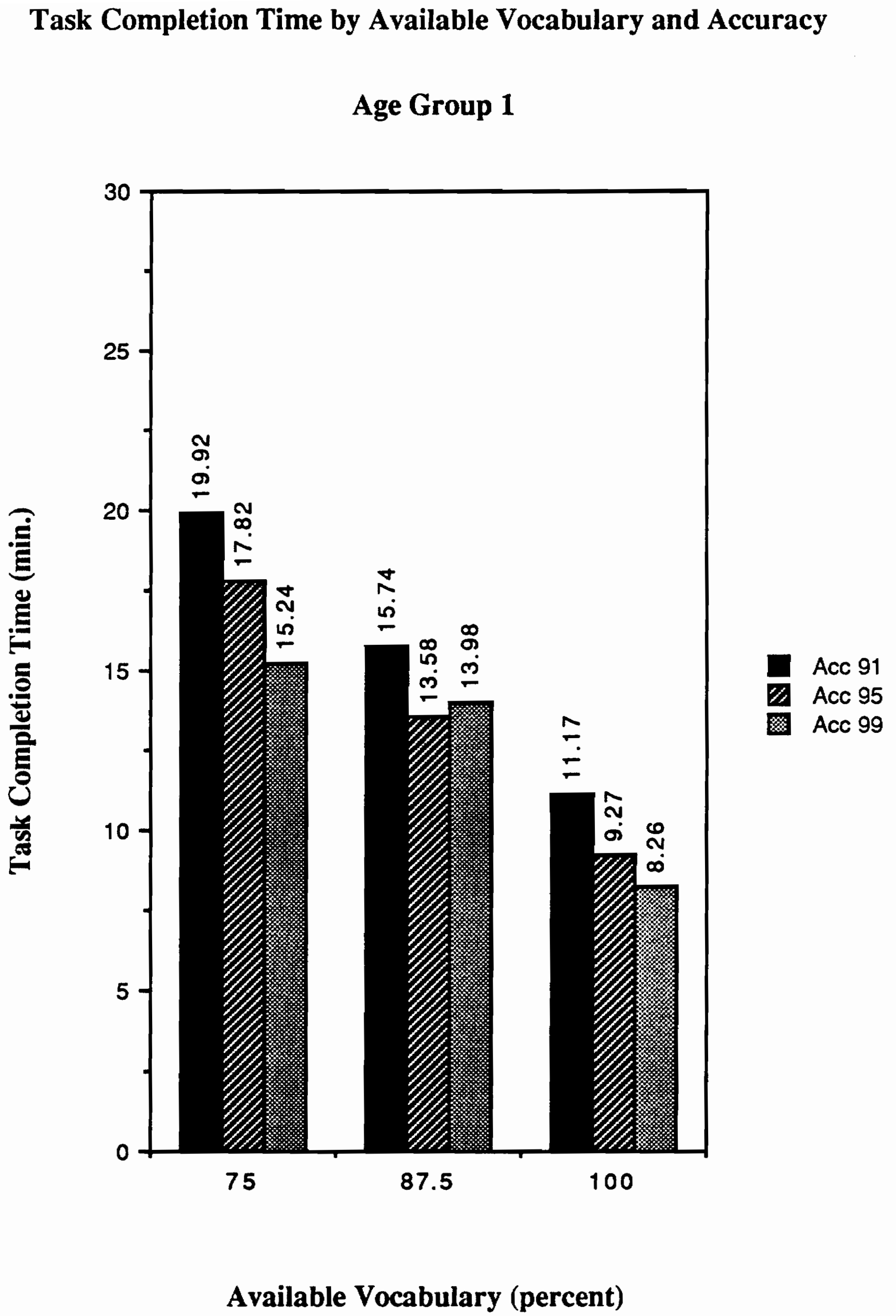


Figure 5. Task Completion Time by Available Vocabulary and Accuracy (Age Group 1)

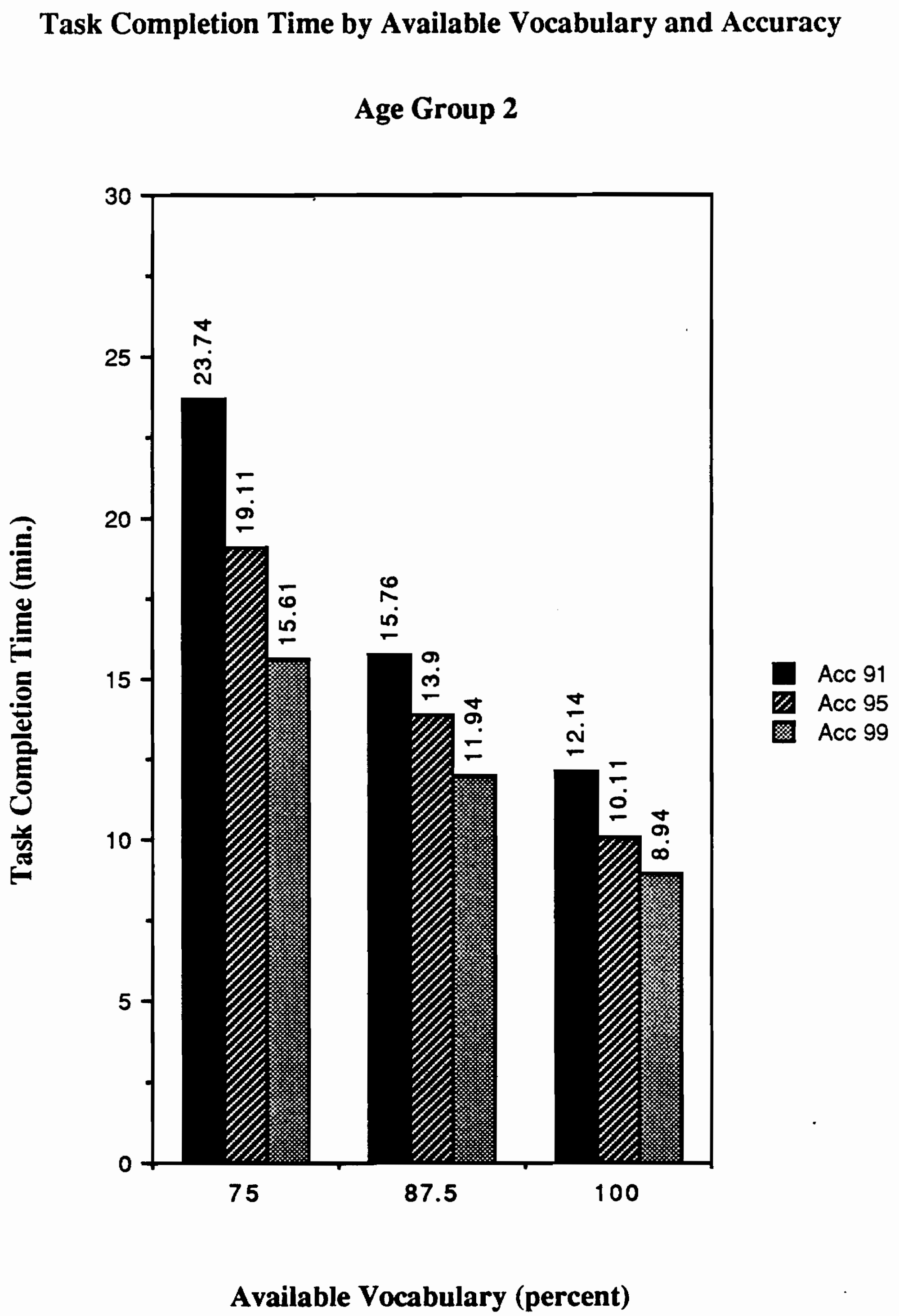


Figure 6. Task Completion Time by Available Vocabulary and Accuracy (Age Group 2)

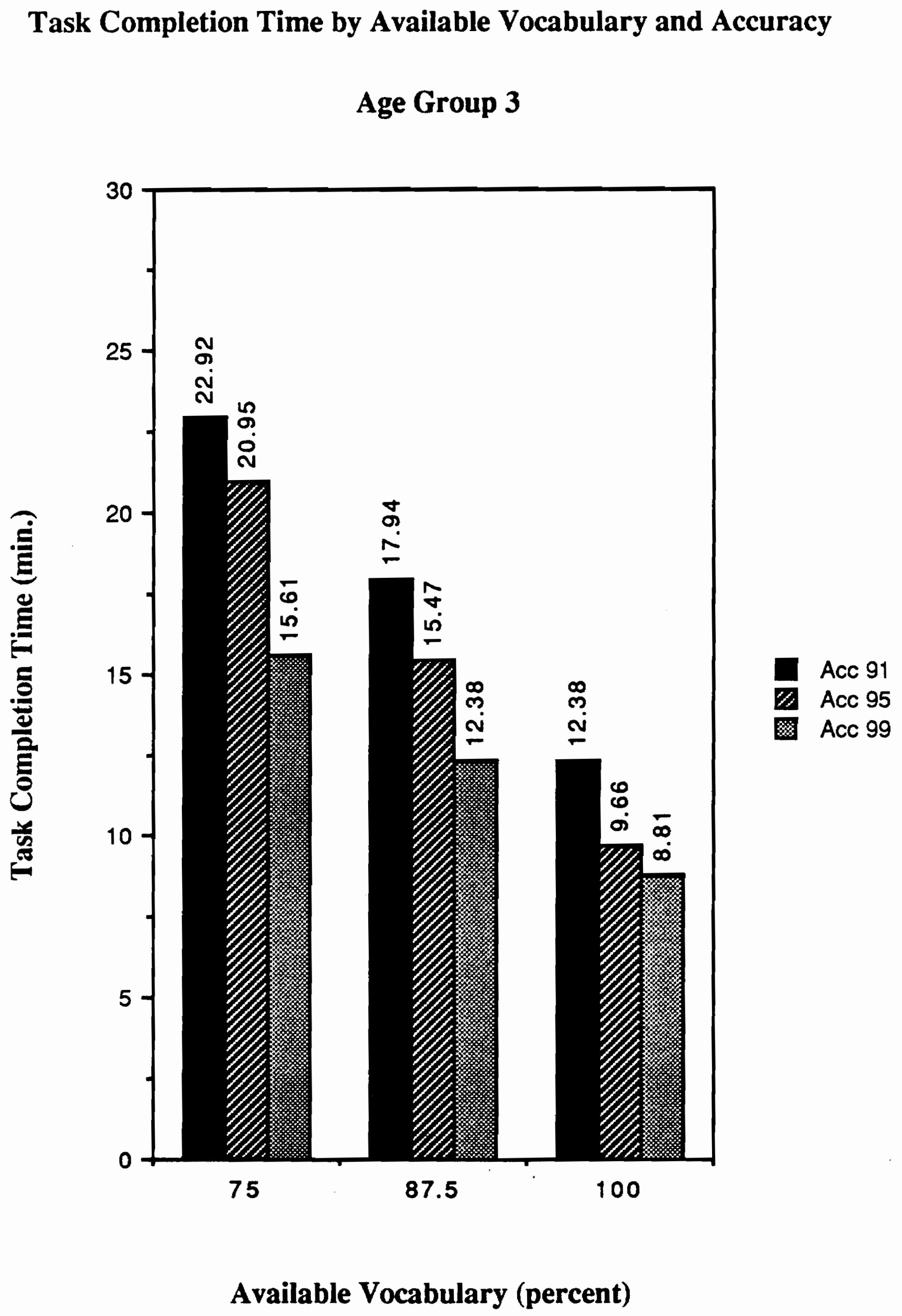


Figure 7. Task Completion Time by Available Vocabulary and Accuracy (Age Group 3)

Table 6. Anova for Task Completion Time (Age Group 1)

Source	*df*	*SS*	*F*	*Prob*
Between Subjects				
Subjects (S)	5	19226.37		
Within Subject				
Acc	2	321034.04	28.13	0.0001
Acc x S	10	57060.41		
Voc	2	2159708.48	31.40	0.0001
Voc x S	10	343864.63		
Acc x Voc	4	63873.63	0.92	0.4714
Acc x Voc x S	20	346883.26		
Total	53	3311690.81		

Table 7. Anova for Task Completion Time (Age Group 2)

Source	*df*	*SS*	*F*	*Prob*
Between Subjects				
Subjects (S)	5	190172.39		
Within Subject				
Acc	2	831068.44	83.31	0.0001
Acc x S	10	49880.67		
Voc	2	2727090.44	410.91	0.0001
Voc x S	10	33183.67		
Acc x Voc	4	158289.78	11.25	0.0001
Acc x Voc x S	20	70368.44		
Total	53	4060056.83		

Table 8. Anova for Task Completion Time (Age Group 3)

Source	*df*	*SS*	*F*	*Prob*
Between Subjects				
Subjects (S)	5	911453.33		
Within Subject				
Acc	2	977488.00	40.76	0.0001
Acc x S	10	119914.67		
Voc	2	2949366.78	42.01	0.0001
Voc x S	10	350991.22		
Acc x Voc	4	124871.56	3.10	0.0389
Acc x Voc x S	20	201619.78		
Total	53	5635705.33		

Table 9. Student-Newman-Keuls Test for Task Completion Time (Age Group 1)

Voc	Mean		Acc	Mean	
75	1059.56	A	91	935.50	A
87.5	865.94	B	95	813.39	B
100	573.06	C	99	749.67	C

*Means with the same letter are not significantly different from each other at $p <= 0.05$

Table 10. Student-Newman-Keuls Test for Task Completion Time (Age Group 2)

Voc	Mean	Acc	Mean	Acc x Voc	Mean
75	1169.17 A	91	1032.72 A	91 x 75	1424.17 A
87.5	831.89 B	95	862.50 B	95 x 75	1146.83 B
100	623.78 C	99	729.61 C	91 x 87.5	945.67 C
				99 x 75	936.50 C
				95 x 87.5	833.83 D
				91 x 100	728.33 E
				99 x 87.5	716.17 E
				95 x 100	606.83 F
				99 x 100	536.17 F

*Means with the same letter are not significantly different from one another at p<= 0.05

Table 11. Student-Newman-Keuls Test for Task Completion Time (Age Group 3)

Voc	Mean		Acc	Mean		Acc x Voc	Mean	
75	1189.44	A	91	1064.78	A	91 x 75	1375.00	A
87.5	915.72	B	95	921.44	B	95 x 75	1256.67	A
100	617.17	C	99	736.11	C	91 x 87.5	1076.33	B
						99 x 75	936.67	C
						95 x 87.5	928.00	C
						91 x 100	743.00	D
						99 x 87.5	742.83	D
						95 x 100	579.67	E
						99 x 100	528.83	E

*Means with the same letter are not significantly different from one another at $p <= 0.05$

crossproduct terms. The procedure, however, uses a pooled error term including the between-subjects variation and the lack of fit in all significance testing. Therefore, the error terms in all regressions have been adjusted to remove systematic variance due to the subjects effects and the lack of fit, and the significance of the linear, quadratic, and crossproduct terms as well as lack of fit have been adjusted accordingly.

Using RSREG, regression equations were generated first for the task completion time data of all age groups combined, and then for each age group separately. In both cases, the coded values for Accuracy and Available vocabulary in the regression equations were -1, 0, and 1 for 91, 95, and 99 percent and 75, 87.5, and 100 percent respectively. These values insure that the design is orthogonal therefore allowing one to compare directly the standardized partial regression weights (beta values) in determining the relative contribution of each parameter in the regression equations. The regressions for task completion time are shown in Tables 12, 13, 14, and 15.

Three dimensional plots of the regression equations using the uncoded variables are shown in Figures 8, 9, 10, and 11. It is apparent from the plots that the regression equations of each separate age group closely resemble that of the combined groups. In each case, examination of the beta weights confirm that decreasing the available vocabulary level and/or the accuracy level increases the task completion time. However, subtle differences do exist between the three age groups. Most notably, when a speech recognizer with greater capability (i.e. a large available vocabulary and/or high recognition accuracy) was simulated, all age groups performed similarly, as is illustrated by the similarity between the forward-most portion of each of the three curves. However, the rear-portions of the curves differ more notably. At lower

Table 12. RSREG for Task Completion Time (Combined Age Groups)

Model	*df*	*SS*	*F*	*Prob*
Linear	2	118.9958	695.5970	0.001
Quadratic	2	0.0356	0.1041	
Crossproduct	1	2.4217	14.1561	0.001
Total Reg.	5	121.4532	141.9908	0.001
Residual				
Subjects	17	15.5628	5.3516	0.001
Lack of Fit	3	0.7183	1.3996	
Adj. Error	136	23.2657		
Total Resid.	156	39.5468		

Parameter	*df*	*Estimate*	*St.Dev*	*T ratio*	*Prob*
Intercept	1	-0.0228	0.0885	-0.26	
Acc	1	-0.4767	0.0484	-9.84	0.001
Voc	1	-0.9352	0.0484	-19.30	0.001
Acc x Acc	1	0.0313	0.0839	0.37	
Voc x Voc	1	0.0029	0.0839	0.04	
Acc x Voc	1	0.1834	0.0593	3.09	0.001

Table 13. RSREG for Task Completion Time (Age Group 1)

Model	*df*	*SS*	*F*	*Prob*
Linear	2	29.8660	65.2837	0.001
Quadratic	2	0.4869	1.0643	
Crossproduct	1	0.2201	0.9620	
Total Reg.	5	30.5730	26.7352	0.001
Residual				
Subjects	5	0.2357	0.2059	
Lack of Fit	3	0.5615	0.8181	
Adj. Error	40	9.1497		
Total Resid.	48			

Parameter	*df*	*Estimate*	*St.Dev*	*T ratio*	*Prob*
Intercept	1	-0.0884	0.1385	-0.64	
Acc	1	-0.3250	0.0759	-4.28	0.001
Voc	1	-0.8509	0.0759	-11.21	0.001
Acc x Acc	1	0.1021	0.1314	0.78	
Voc x Voc	1	-0.1736	0.1314	-1.32	
Acc x Voc	1	0.0958	0.0929	1.03	

Table 14. RSREG for Task Completion Time (Age Group 2)

Model	*df*	*SS*	*F*	*Prob*
Linear	2	42.8721	456.7421	0.001
Quadratic	2	0.6636	14.1393	0.001
Crossproduct	1	1.6026	34.1465	0.001
Total Reg.	5	45.1382	192.35	0.001
Residual				
Subjects	5	2.3268	9.9163	0.001
Lack of Fit	3	0.3341	2.3732	
Adj. Error	40	1.8773		
Total Resid.	48	4.5383		

Parameter	*df*	*Estimate*	*St.Dev*	*T ratio*	*Prob*
Intercept	1	-0.1829	0.0936	-1.96	0.01
Acc	1	-0.5301	0.0512	-10.34	0.001
Voc	1	-0.9539	0.0512	-18.61	0.001
Acc x Acc	1	0.0653	0.0888	0.74	
Voc x Voc	1	0.2259	0.0888	2.55	0.001
Acc x Voc	1	0.2584	0.0676	4.12	0.001

Table 15. RSREG for Task Completion Time (Age Group 3)

Model	*df*	*SS*	*F*	*Prob*
Linear	2	47.9594	116.5713	0.001
Quadratic	2	0.0874	0.2121	
Crossproduct	1	0.9223	4.4828	0.05
Total Reg	5	48.9690	47.6080	0.001
Residual				
Subjects	5	11.1520	10.8421	0.001
Lack of Fit	3	0.6056	0.9813	
Adj. Error	40	1.8773		
Total Resid.	48	19.9863		

Parameter	*df*	*Estimate*	*St.Dev*	*T ratio*	*Prob*
Intercept	1	0.2028	0.1963	1.03	
Acc	1	-0.5748	0.1075	-5.34	0.001
Voc	1	-1.0001	0.1075	-9.31	0.001
Acc x Acc	1	-0.0734	0.1863	-0.39	
Voc x Voc	1	-0.0434	0.1863	-0.23	
Acc x Voc	1	0.1960	0.1317	1.49	0.05

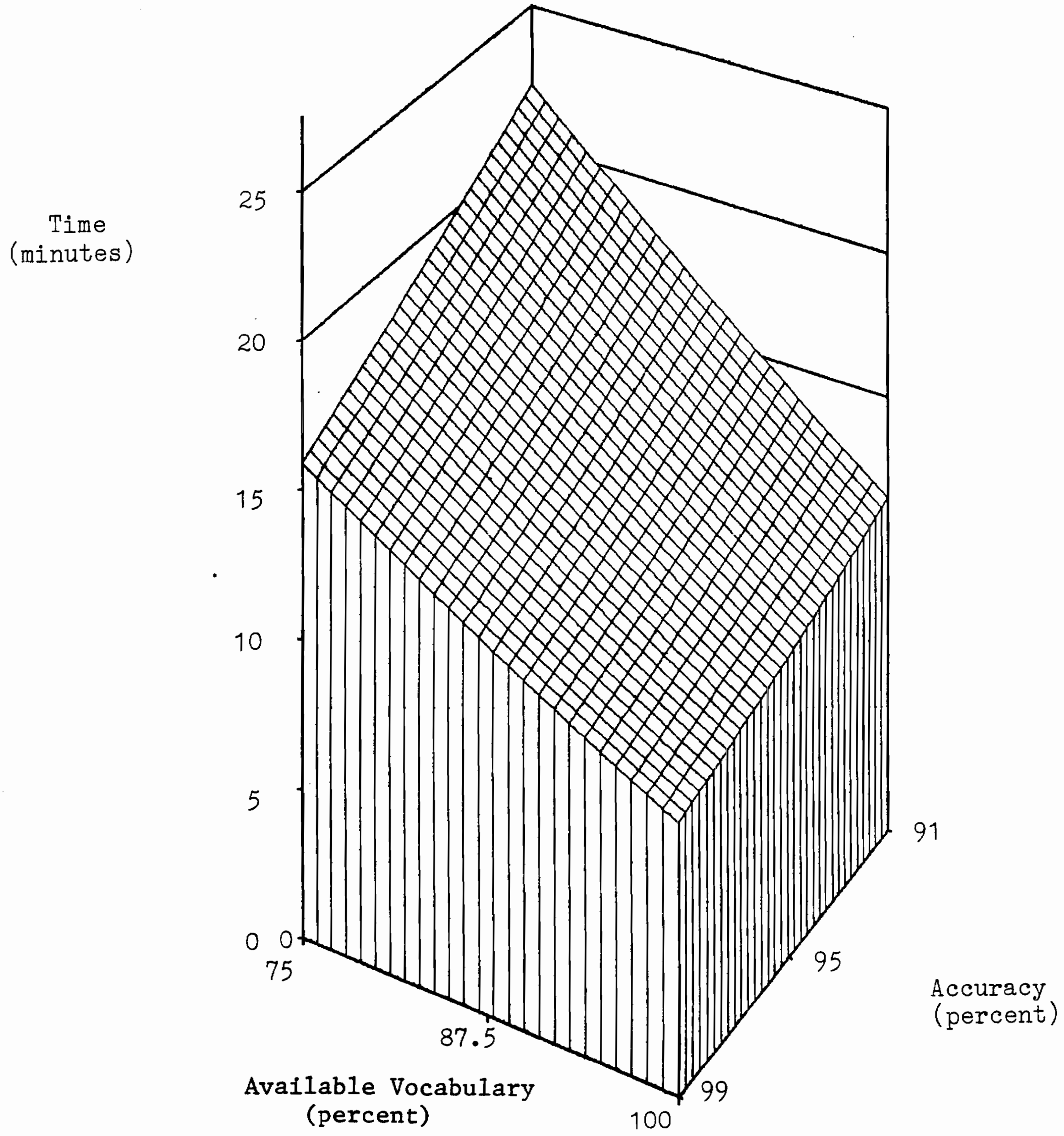


Figure 8. Task Completion Time Response Surface (Combined Age Groups)

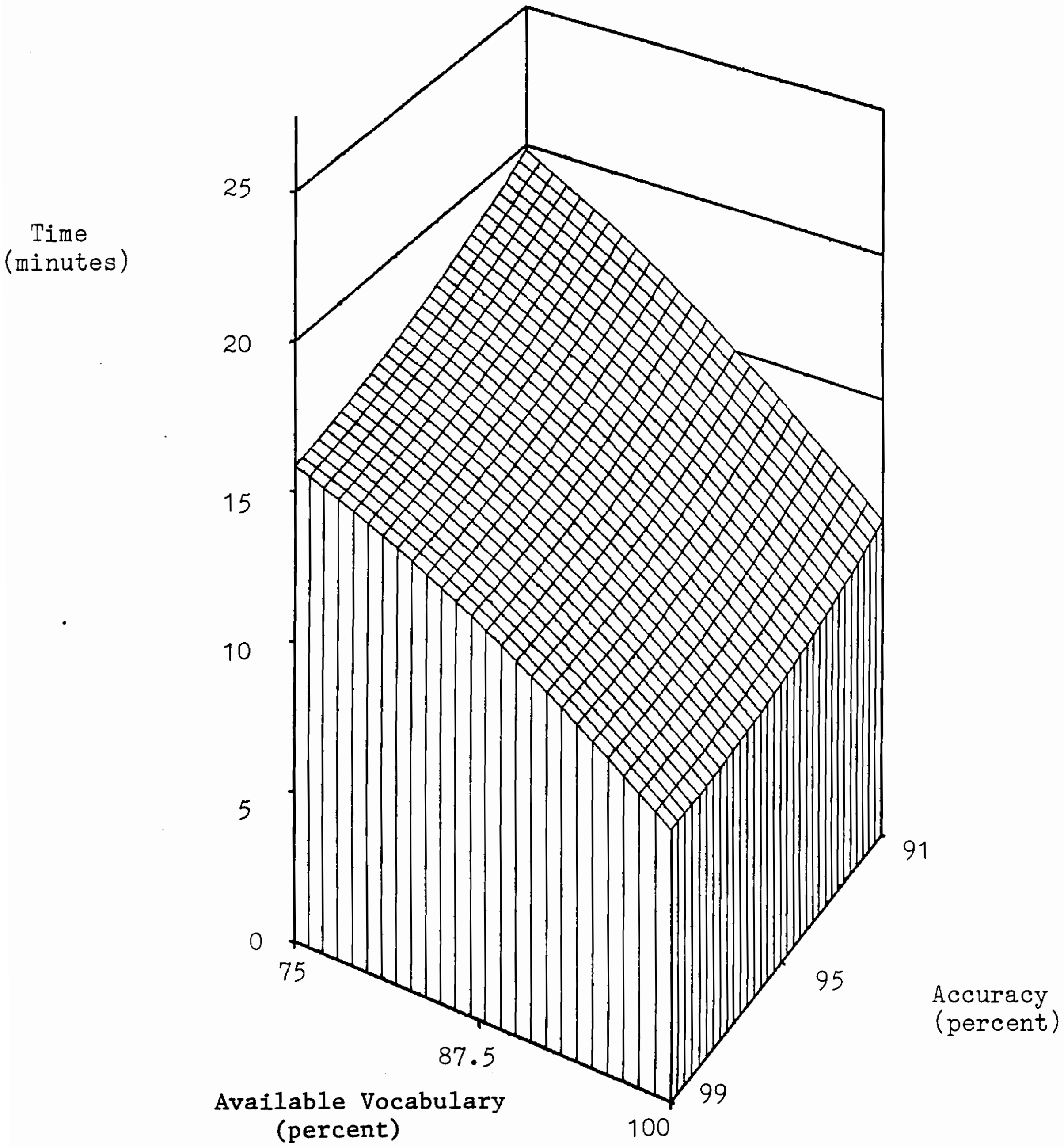


Figure 9. Task Completion Time Response Surface (Age Group 1)

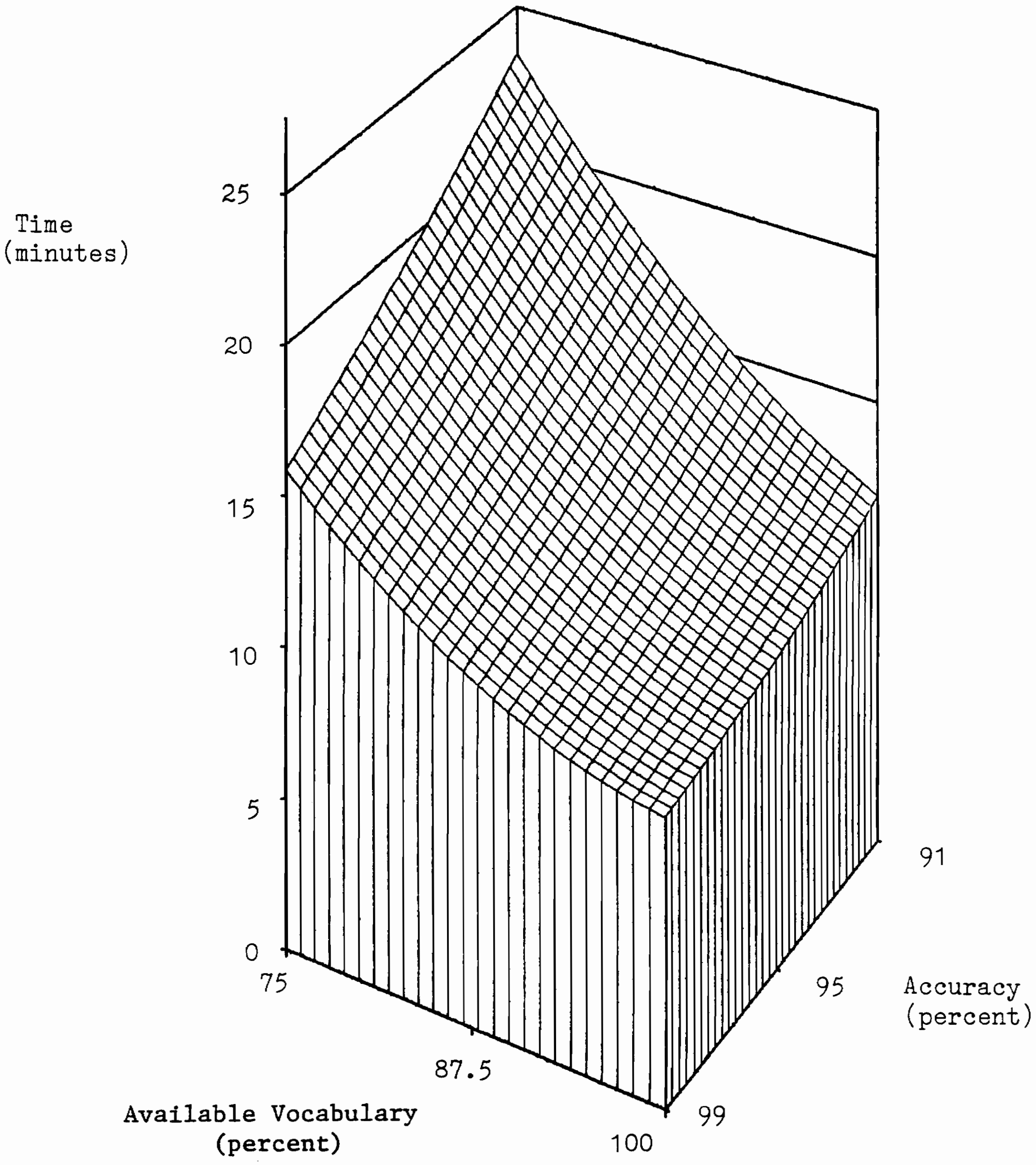


Figure 10. Task Completion Time Response Surface (Age Group 2)

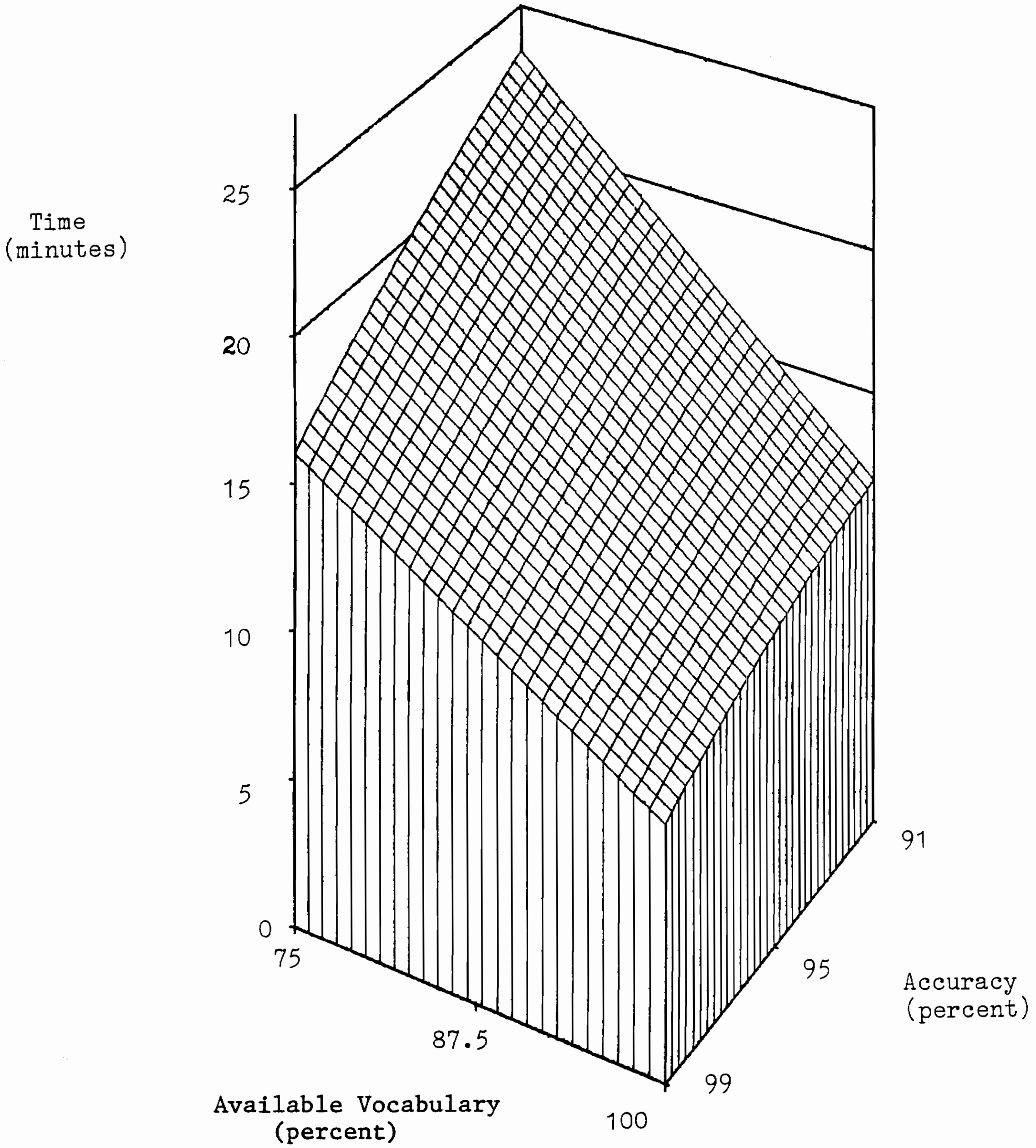


Figure 11. Task Completion Time Response Surface (Age Group 3)

levels of either accuracy or available vocabulary, task completion time increased more rapidly and to a greater degree for the older two groups than for the youngest group as illustrated by the height of the rear peak of each curve.

In none of the four analyses was a significant lack of fit found ($p > 0.05$) suggesting that a complete second-order equation was adequate to account for the variables sampled. In fact, each of the four curves appears nearly linear as indicated by the lack of significance of the quadratic terms ($p > 0.05$) in all but the regression involving age group 2. Even here, only one of the two quadratic terms was found significant. In addition, the subject effect for the models involving age group 2 and age group 3 was significant ($p < 0.001$), but not for the model involving age group 1.

Final Error

Again, Final Error is the percentage of incorrect words (either uncorrected misrecognized words or uncorrected XXXX's) remaining at the completion of each of the 9 experimental trials. An ANOVA was run on the resulting sums using Accuracy, Available Vocabulary, and Age as the main effects. These results are shown in Table 16. The main effects of Accuracy and Available Vocabulary were found significant ($p < 0.001$ and $p < 0.05$ respectively). A Student-Newman-Keuls test was performed on the means of the significant effects as shown in Table 17. Available vocabularies of 75 and 100 percent were significantly different from one another but neither was different from the 87.5 percent vocabulary condition. Tasks involving the 99 percent

Table 16. Anova for Final Error Rates

Source	*df*	*SS*	*F*	*Prob*
Between Subjects				
Age	2	5.78	0.17	0.8478
Subject (S I Age)	15	259.61		
Within Subject				
Acc	2	32.93	10.74	0.0003
Acc x Age	4	11.29	1.84	0.1468
Acc x S(Age)	30	46.00		
Voc	2	19.00	3.72	0.0361
Voc x Age	4	17.22	1.68	0.1795
Voc x S(Age)	30	76.67		
Acc x Voc	4	3.18	0.56	0.6937
Acc x Voc x Age	8	12.37	1.08	0.3864
Acc x Voc x S(Age)	60	85.55		
Total	161	569.61		

Table 17. Student-Newman-Keuls Test for Final Error Rates

Voc	Mean		Acc	Mean	
75	1.8148	A	91	1.9259	A
87.5	1.4815	A B	95	1.5185	A
100	0.9815	B	99	0.8333	B

*Means with the same letter are not significantly different from each other at $p <= 0.05$

accuracy level resulted in significantly fewer remaining errors than those involving either the 91 or 95 percent accuracy levels.

Figure 12 shows the mean final error rate by accuracy and available vocabulary. Although the maximum error rate on a single trial was nearly 5 percent (occurring at a vocabulary level of 75 percent and an accuracy level of 95 percent) the mean rates range from 0.67 percent to 2.39 percent.

A quadratic regression was performed on the coded final error data using RSREG. Because neither the main effect of Age nor any interactions involving age were found to be significant ($p > 0.05$), a single empirical model was developed describing Final Error in terms of Available vocabulary and Accuracy for all age groups combined. As before, the error term has been refined to insure proper significance testing, taking into account systematic variation due to lack of fit and subjects. The regression is shown in Table 18. A three dimensional plot is shown in Figure 13.

Only the linear effects of Accuracy and Available vocabulary were significant. Examination of the beta weights confirm that decreasing either the accuracy level or the percent of task vocabulary recognized produces an increase in the average number of uncorrected errors, and that the two variables contribute nearly equally to the regression equation.

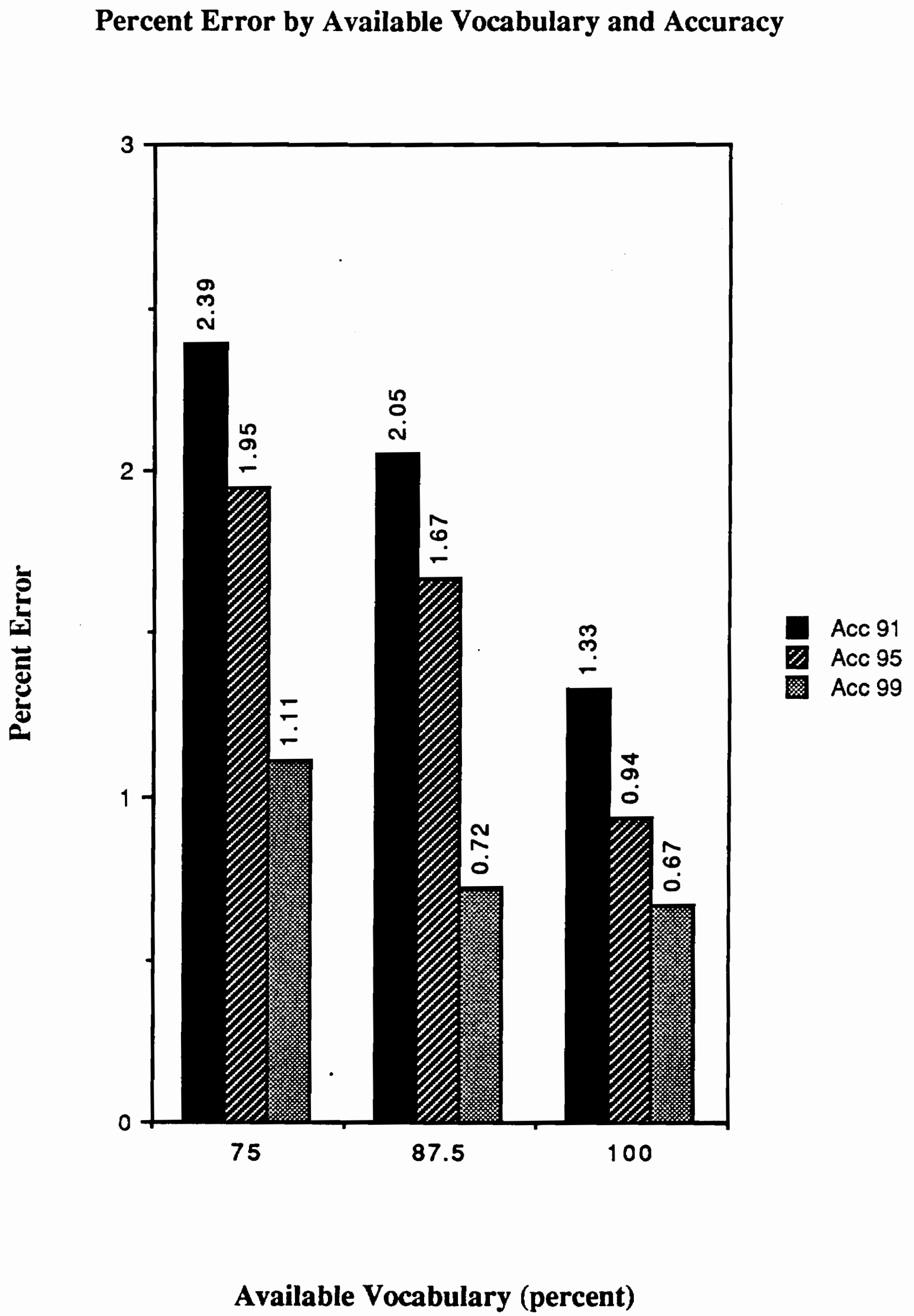


Figure 12. Percent Final Error by Available Vocabulary and Accuracy

Table 18. RSREG for Final Error Rates

Model	*df*	*SS*	*F*	*Prob*
Linear	2	14.4099	13.9166	0.001
Quadratic	2	0.2669	0.2578	
Crossproduct	1	0.4750	0.9175	
Total Reg	5	15.1518	58.5327	0.001
Residual				
Subjects	17	75.0119	8.5228	0.001
Lack of Fit	3	0.4253		
Adj. Error	136	70.4109		
Total Resid.	156	145.8482		

Parameter	*df*	*Estimate*	*St.Dev*	*T ratio*	*Prob*
Intercept	1	0.0788	0.1699	0.46	
Acc	1	-0.2904	0.0930	-3.12	0.001
Voc	1	-0.2215	0.0930	-2.38	0.005
Acc x Acc	1	-0.0738	0.1612	-0.46	
Voc x Voc	1 .	-0.0443	0.1612	-0.27	
Acc x Voc	1	0.0812	0.1139	0.71	

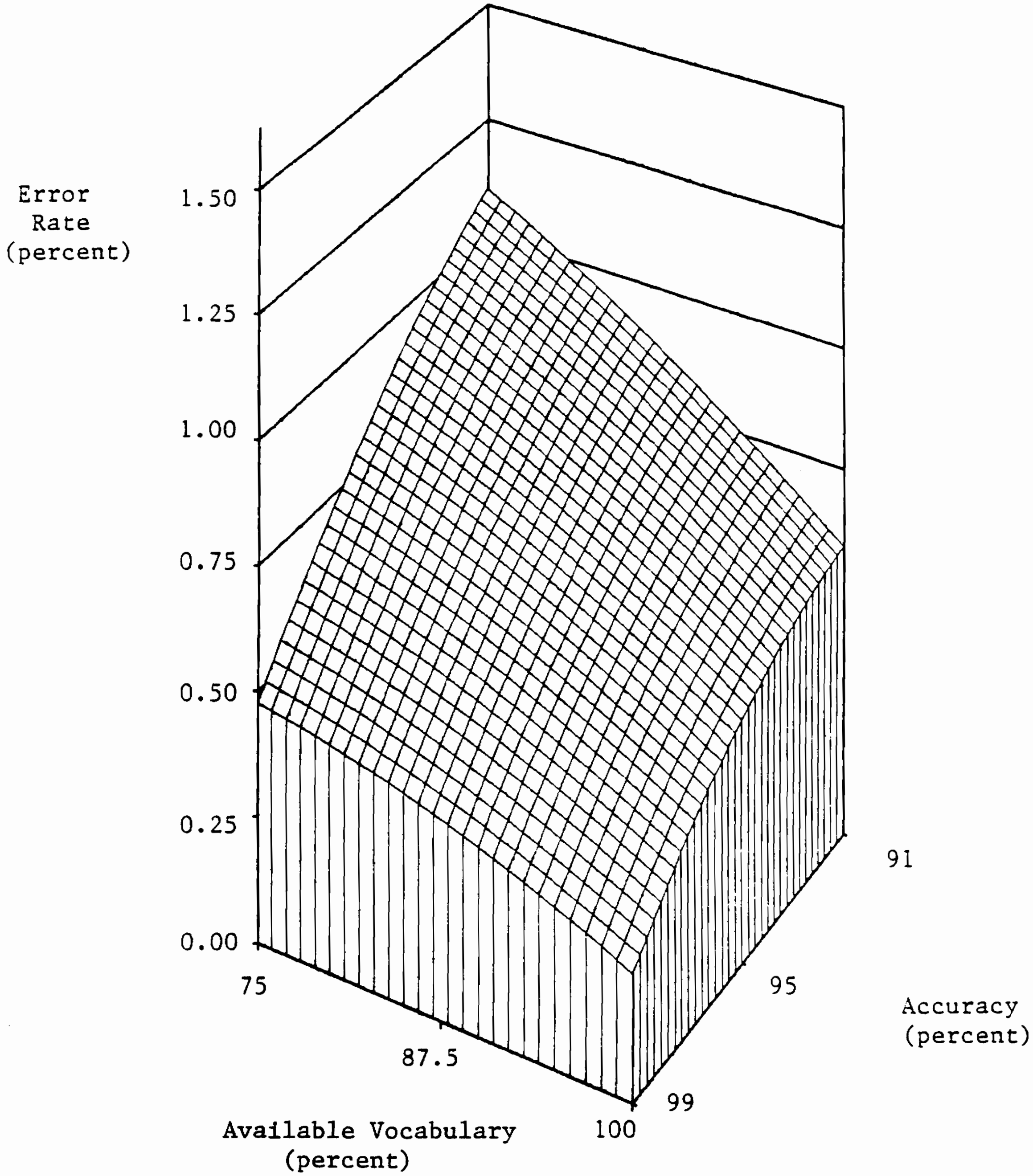


Figure 13. Percent Final Error Response Surface

Subjective Rating

As described previously, following each treatment condition, each subject responded to a set of 13 bipolar rating scales. Each scale has seven possible intervals which may have been chosen. Each of the intervals was assigned a value ranging from one to seven. A high value means that the subject perceived the voice input system in a positive fashion (e.g. fast, pleasing, facilitating, etc.), while a low value means the subject perceived the system in a negative, or detrimental sense (e.g. slow, irritating, etc.).

Because the primary purpose of this analysis was to determine the effects of accuracy and available vocabulary on the subject's judgement of whether the voice input system is an acceptable input device, a single metric, "acceptability index" (AI) was first developed. Not only did this result in a single number measure of acceptability to be used in all subsequent analyses, but it also indicates which of the adjectives used in the bipolar scales are truly attributes of acceptability.

Since "acceptability" is of primary interest, a Spearman Rank correlation coefficient was computed for each bipolar scale with the acceptable/unacceptable scale. Those adjective scales achieving high correlation were considered attributes of acceptability and were therefore used to form the AI. As shown in Table 19, each of the 12 variables resulted in a relatively large correlation coefficient when compared with the acceptability scale. The "simple/complicated" scale had the lowest correlation coefficient (0.5942) yet was still significantly correlated ($p < 0.001$). As such, the acceptability index (AI) was developed to include each of the individual adjective

Table 19. Spearman Correlation Coefficients of Bipolar Scales with Acceptability

Bipolar Scale	Rho
Fast / Slow	0.7034
Accurate / Inaccurate	0.6874
Natural / Unnatural	0.7261
Complete / Incomplete	0.7718
Comfortable/ Uncomfortable	0.7618
Consistent / Inconsistent	0.8285
Pleasant / Unpleasant	0.8366
Dependable / Undependable	0.8940
Friendly / Unfriendly	0.8940
Facilitating / Distracting	0.7728
Simple / Complicated	0.5942
Useful / Useless	0.8489

scales. This measure was computed as the sum of the 12 individual scale response values for each subject under each treatment condition. Therefore, the AI measure had a range of 12 to 84, with higher values indicating a more acceptable system.

Because bipolar rating scales can not necessarily be considered as providing interval scale data, all analyses of the subjective rating scale data employed only nonparametric statistical tests.

The AI data were first subjected to a Kruskal-Wallis one-way analysis of variance using Age as the variable of interest. As shown in Table 20, there were statistically significant differences among age groups as measured by the AI, using a chi-square large sample approximation ($p < 0.001$). Post-hoc pair-wise comparisons were performed using the Kolmogorov-Smirnov two-sample test in order to locate the significance of the age main effect. To ensure protection against Type I error, alpha was set equal to 0.0167. As shown in Table 21, results indicate age groups 1 and 2 did not differ statistically from on another as measured by the AI ($p > 0.05$), however both differed from age group 3 ($p < 0.005$). As such, subsequent analyses combined ages 1 and 2 to form Group 1 and subjects of age group 3 will be referred to as Group 2. In addition to the statistical reasoning behind this grouping, there is a practical reason to support it, which will be discussed in the discussion section.

A Friedman two-way analysis of variance was performed on the AI data of each group, first using Accuracy as the variable of interest and then using Available vocabulary. As Table 22 indicates, the main effect of Accuracy was significant for both age groups ($p < 0.001$). Post-hoc pair-wise comparisons using the Wilcoxon matched-pairs signed-ranks test were performed (again, using a conservative alpha level) and the results are shown in Table 23. For group 1, all three accuracy levels

Table 20. Kruskal-Wallis 1-Way Anova on AI for the Main Effect of Age

N=162 Chi-square=21.6486 p<0.001

Mean Rank	Age Group
69.39	1
69.39	2
105.72	3

Table 21. Kolmogorov-Smirnow 2-Sample Test on AI for Age

$N_1 = N_2 = N_3 = 54$ Alpha = 0.0167

Age Group	
1 (20-25)	A
2 (35-40)	A
3 (50-55)	B

Table 22. Friedman 2-Way Anova on AI for the Accuracy Main Effect

K = 3 Alpha = 0.05

Group 1

N = 38 Chi-square = 22.38 p < 0.0001

Accuracy	Mean Rank
91	1.47
95	1.94
99	2.58

Group 2

N = 18 Chi-square = 16.028 p < 0.0003

Accuracy	Mean Rank
91	1.47
95	1.78
99	2.75

Table 23. Wilcoxon Matched-Pairs Signed-Ranks Tests on AI for Accuracy

Group 1		Group 2	
Acc		Acc	
91	A	91	A
95	B	95	A
99	C	99	B

* Means with the same letter are not significantly different at $p < 0.0167$.

were significantly different from one another with the acceptability rating decreasing as the accuracy level decreases. Group 2 rated the 99 percent accuracy conditions as significantly more acceptable than either of the other two levels.

The results of the Friedman two-way analysis of variance using Available vocabulary as the main effect are shown in Table 24. The mean ranks indicate that the acceptability rating increases as the vocabulary level increases, however these differences were only significant for Group 2. Post-hoc comparisons using the Wilcoxon sign test failed to locate any significant differences between the pairs when the alpha level was set to ensure protection against Type I error. These results are shown in Table 25.

In order to determine if the interaction of Accuracy and Available vocabulary is significant, the nine treatment conditions were coded as a dummy variable containing 9 levels and were subjected to the Friedman two-way ANOVA. As Table 26 indicates, the interaction is significant ($p < 0.01$) for both Groups 1 and 2. As shown in Table 27, post-hoc pair-wise comparisons were performed using the Wilcoxon sign rank test. When alpha was adjusted to ensure protection against Type I error ($K=9$, alpha $= 0.0014$), no significant differences were found between pairs for either group.

Figures 14 and 15 show the mean rank assigned each condition in the proceeding analysis for each group. The sharpest decline in user preference occurs in going from a 99 to a 95 percent accuracy condition, and to a slightly lesser extent in going from 95 to 91 percent. The small effect of available vocabulary level on user satisfaction is also apparent in this figure.

A quadratic regression was also performed on the acceptability index data for each group using RSREG. However, in order to facilitate comparison between the three

Table 24. Friedman 2-Way Anova on AI for the Available Vocabulary Main Effect

K = 3 Alpha = 0.05

Group 1

N = 38 Chi-square = 3.792 p < 0.1502

Vocabulary	Mean Rank
75	1.76
87.5	2.01
100	2.22

Group 2

N = 18 Chi-square = 7.750 p < 0.0003

Vocabulary	Mean Rank
75	1.50
87.5	2.08
100	2.42

Table 25. Wilcoxon Matched-Pairs Signed-Ranks Test on AI for Available Vocabulary

Group 1		Group 2	
Vocabulary		Vocabulary	
75	A	75	A
87.5	A	87.5	A
100	A	100	A

* Means with the same letter are not significantly different at $p < 0.0167$.

Table 26. Friedman 2-Way Anova on AI for the Accuracy x Available Vocabulary Interaction

K=3 Alpha=0.05

Group 1

N=12 Chi-square=20.722 p < 0.0079

Condition (Acc. Voc)		Mean Rank
4	(91, 87.5)	3.25
2	(95, 75)	3.88
1	(91, 75)	3.92
7	(91, 100)	3.96
8	(95, 100)	5.38
5	(95, 87.5)	5.46
6	(99, 87.5)	6.08
3	(99, 75)	6.33
9	(99, 100)	6.75

Group 2

N=6 Chi-square=24.800 p < 0.0017

Condition (Acc. Voc)		Mean Rank
1	(91, 75)	2.17
4	(91, 87.5)	3.50
2	(95, 75)	3.75
5	(95, 87.5)	3.92
7	(91, 100)	4.25
8	(95, 100)	5.67
3	(99, 75)	6.17
6	(99, 87.5)	7.67
9	(99, 100)	7.92

Table 27. Wilcoxon Matched-Pairs Signed-Ranks Tests on AI for Accuracy x Available Vocabulary Interaction

Group 1

Condition	(Acc. Voc)	
4	(91, 87.5)	A
2	(95, 75)	A
1	(91, 75)	A
7	(91, 100)	A
8	(95, 100)	A
5	(95, 87.5)	A
6	(99, 87.5)	A
3	(99, 75)	A
9	(99, 100)	A

Group 2

Condition	(Acc. Voc)	
1	(91, 75)	A
4	(91, 87.5)	A
2	(95, 75)	A
5	(95, 87.5)	A
7	(91, 100)	A
8	(95, 100)	A
3	(99, 75)	A
6	(99, 87.5)	A
9	(99, 100)	A

* Means with the same letter are not significantly different at p < 0.0014.

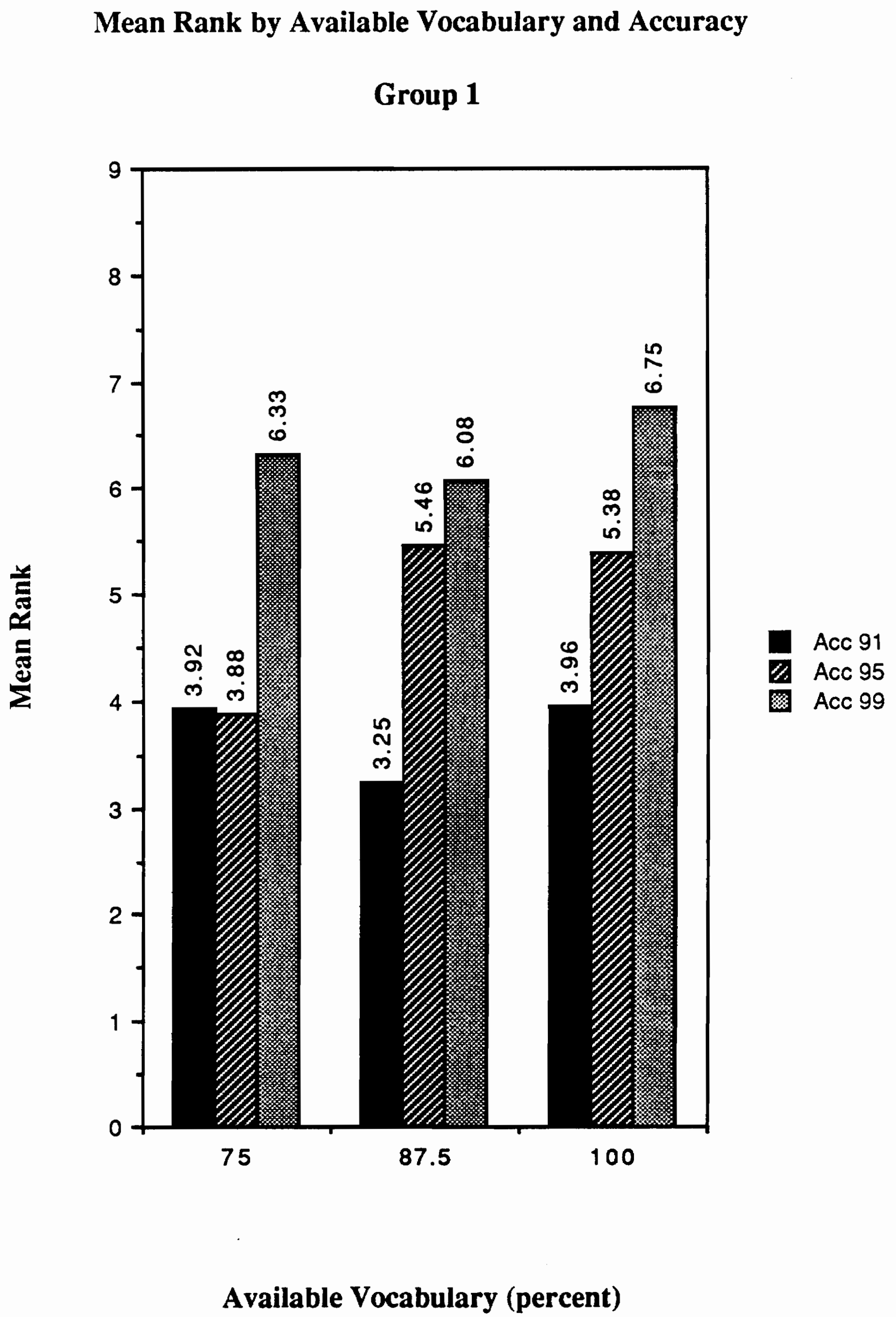


Figure 14. INVAI by Available Vocabulary and Accuracy (Group 1)

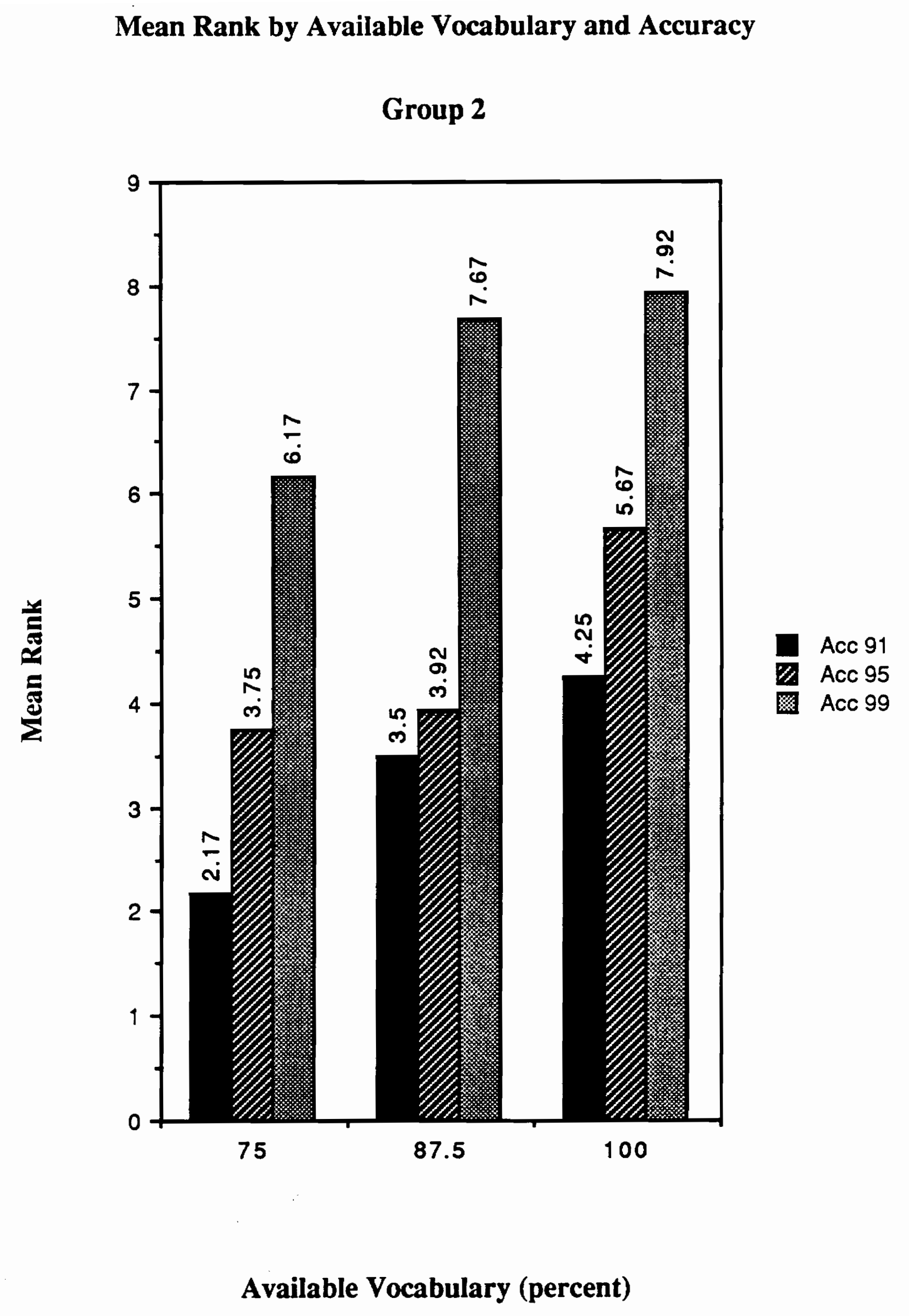


Figure 15. INVAI by Available Vocabulary and Accuracy (Group 2)

dimensional plots of the regressions of the three dependent measures (Task Completion Time, Final Error, and AI) the AI data were first recoded. Because large values of "Time" and "Error" correspond to poor performance and small values correspond to good performance, the AI data were recoded so that large values would corr϶spond to a poor acceptability rating and small values would correspond to a good rating. To avoid confusion, this measure is referred to as the INVAI (inverse acceptability index).

The regressions on the INVAI data for the two groups are shown in Tables 28 and 29. Three dimensional plots of the equations are shown in Figures 16 and 17 respectively. The regressions for the two groups are quite similar. In each case, only the linear effects of Accuracy and Available vocabulary are significant. Examination of the beta weights confirm that decreasing either the accuracy or vocabulary level decreases the acceptability of the system for both age groups. For the two groups, the relative contributions of the two independent variables were nearly the same with Accuracy having approximately twice the effect of Available vocabulary. In general, Group 2 rated the speech recognition system more favorably, regardless of the condition than did Group 1.

Table 28. RSREG for Inverse AI (INVAI) (Group 1)

Model	*df*	*SS*	*F*	*Prob*
Linear	2	17.6492	20.3276	0.001
Quadratic	2	0.4775	0.5499	
Crossproduct	1	0.0034	0.0078	
Total Reg.	5	18.1300	8.3524	0.001
Residual				
Subjects	11	47.1054	9.8642	0.001
Lack of Fit	3	0.2529	0.1942	
Adj. Error	88	38.2031		
Total Resid.	102	85.5614		

Parameter	*df*	*Estimate*	*St.Dev*	*T ratio*	*Prob*
Intercept	1	0.3726	0.1971	1.89	0.05
Acc	1	-0.4481	0.1079	-4.15	0.001
Voc	1	-0.2105	0.1079	-1.95	0.01
Acc x Acc	1	-0.1394	0.1870	-0.75	
Voc x Voc	1	-0.0215	0.1870	-0.12	
Acc x Voc	1	0.0084	0.1322	0.06	

Table 29. RSREG for Inverse AI (INVAI) (Group 2)

Model	*df*	*SS*	*F*	*Prob*
Linear	2	4.1776	11.7874	0.001
Quadratic	2	0.1369	0.3863	
Crossproduct	1	0.0008	0.0043	
Total Reg	5	4.3152	4.8702	0.001
Residual				
Subjects	5	23.0212	25.9818	0.001
Lack of Fit	3	0.0829	0.1560	
Adj. Error	40	7.0882		
Total Resid.	48	30.1924		

Parameter	*df*	*Estimate*	*St.Dev*	*T ratio*	*Prob*
Intercept	1	-0.4632	0.2413	-1.92	0.05
Acc	1	-0.3012	0.1321	-2.28	0.001
Voc	1	-0.1590	0.1321	-1.20	0.05
Acc x Acc	1	-0.1067	0.2289	-0.47	
Voc x Voc	1	0.0056	0.2289	0.02	
Acc x Voc	1	0.0056	0.1619	0.03	

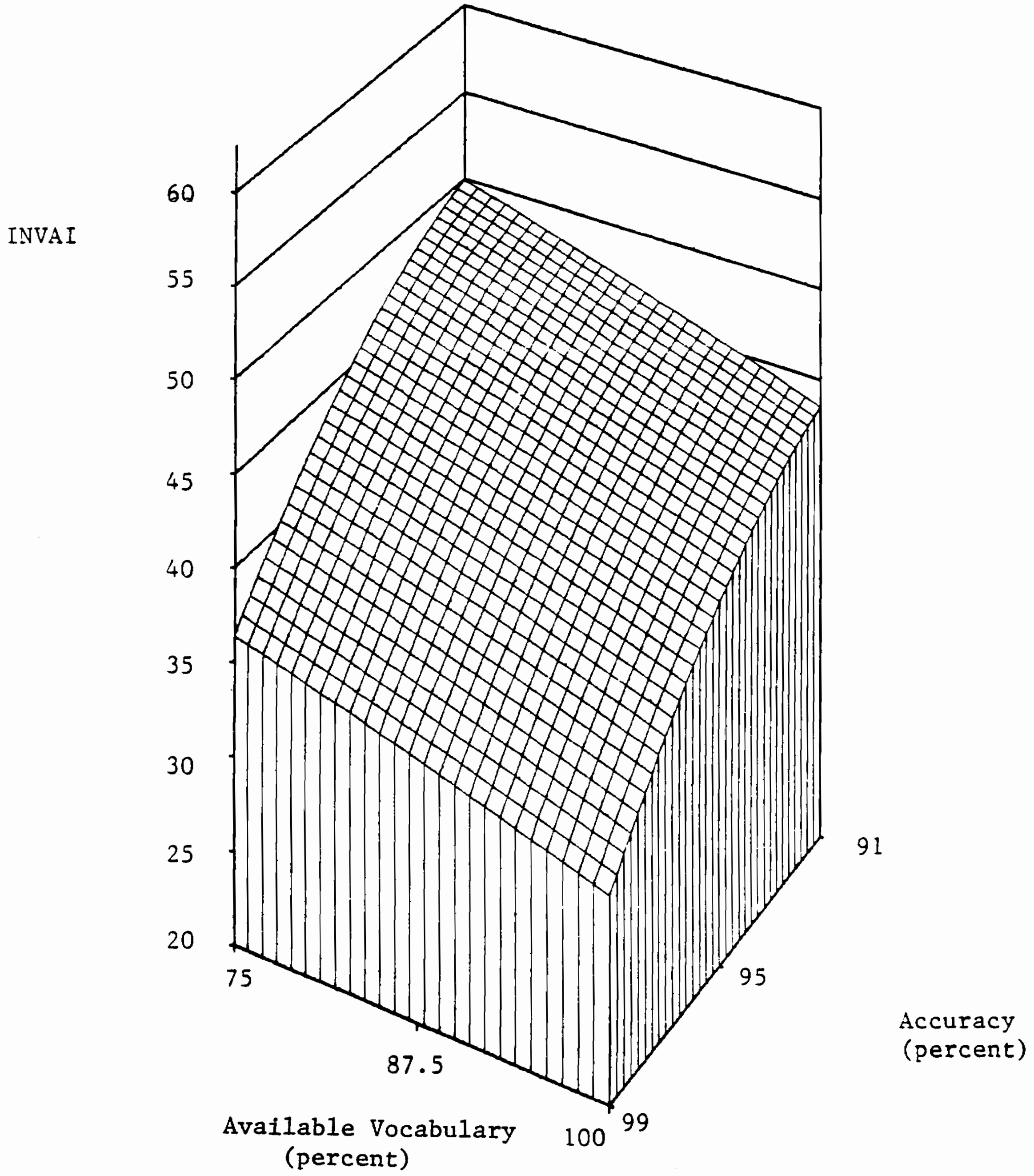


Figure 16. INVAI Response Surface (Group 1)

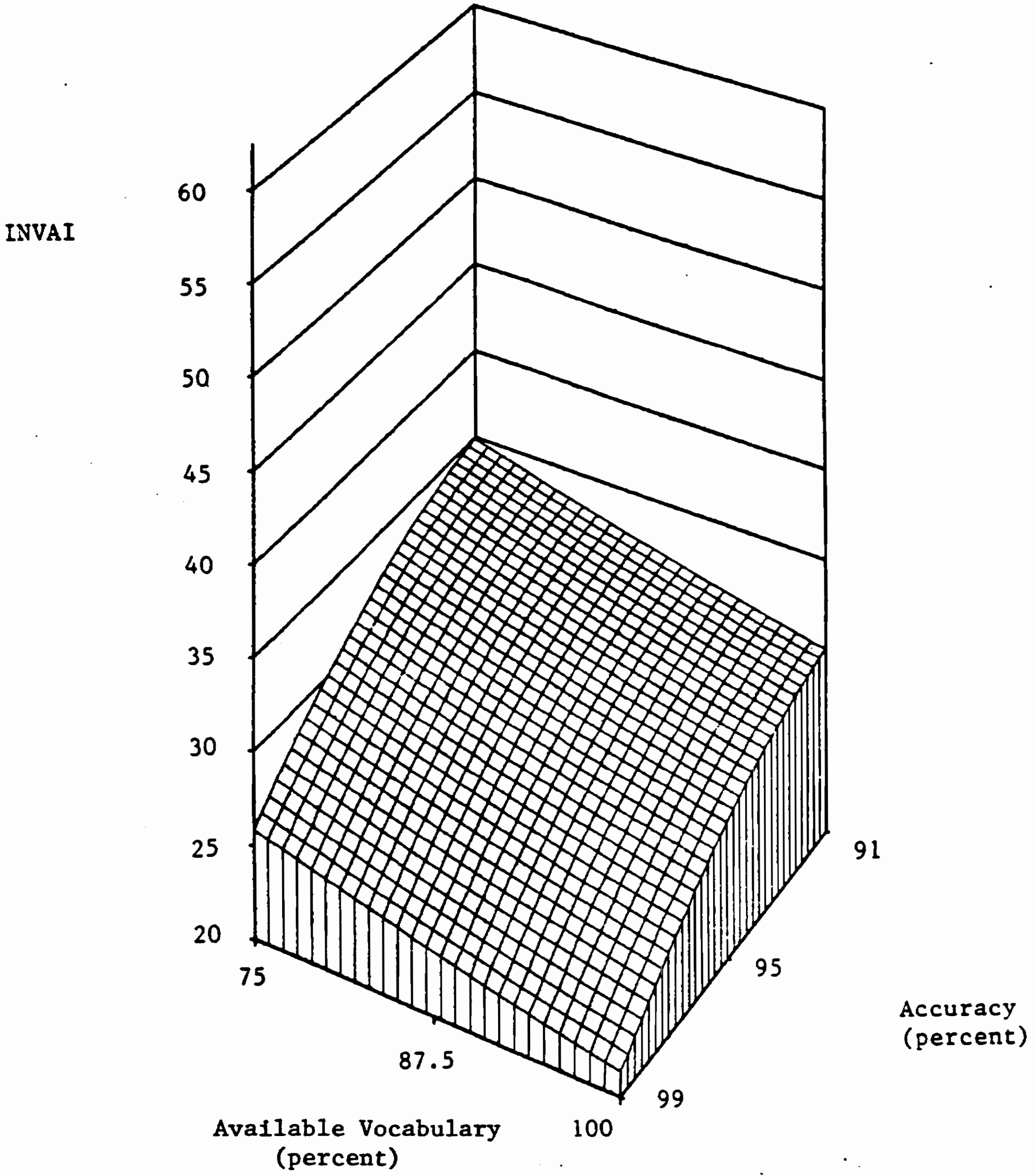


Figure 17. INVAI Response Surface (Group 2)

Discussion

Task Completion Time

The time required to complete the task was significantly affected by the accuracy level as well as the available vocabulary level at which the recognition system was operating. Task completion time increased on average roughly 17 percent in going from the 99 percent accuracy conditions to the 95 percent conditions and approximately 50 percent in going from the 99 percent to the 91 percent. The levels of available vocabulary investigated had a similar but more pronounced effect on task completion time, increasing an average of 44 percent when going from a 100 percent vocabulary to an 87.5 percent vocabulary and nearly 88 percent in going from a 100 percent vocabulary to a 75 percent available vocabulary.

The greater effect of available vocabulary on task completion time is also evident in comparing the beta weights of accuracy and available vocabulary in the regression analyses performed. However, care must be taken in interpreting these results, in that they are entirely dependent upon the ranges of the independent variables, and merely say that task completion time varies more due to a change in available vocabulary of 75 to 100 percent than with a change in accuracy of 91 to 99 percent and *not* that available vocabulary itself is the primary determinant of task completion time. Actually, accuracy level and available vocabulary level together prove to be excellent predictors of task completion time as evidenced by the R-squared values from the regression analyses which range in value from 0.71 to 0.90.

Clearly, even the small differences in accuracy rates as were studied here have a significant impact on the time to complete a data entry task. This suggests that for applications in which information entry rate is of *primary* importance, accuracy levels should be compromised as little as possible. As previously discussed, in order for very high rates to be achieved, the system will likely be expensive, environmental conditions under which the system can be used are limited, and vocabulary sizes will be small, to mention just a few of the compromises involved. However, the results indicate that the corresponding time difference in using a recognizer which operates at 95 rather than 99 percent accuracy is not dramatic. For the task used here, the average difference was roughly 17 percent, which is approximately 10 additional minutes for every hour. Given the difficulty, trade-offs, and expense necessary to gain the higher accuracy level, the time increase associated with the lower accuracy level may be more acceptable in a number of applications.

As future research adds to our body of knowledge concerning the exact relationship between recognizer accuracy and the many factors which affect it, the results of this study will become even more useful. Users can then make knowledgeable choices regarding trade-offs between task completion time and other desirable system features.

As indicated by the nearly 90 percent increase in time needed to complete a task when going from a 100 to a 75 percent available vocabulary, the available vocabulary level also has a major influence on the time needed to complete a task. Just as with accuracy levels, when information entry rate is of primary importance, the largest possible vocabulary size with respect to the task requirements should be used. Input by character-level entry is significantly slower than word level entry and even relatively small decreases in vocabulary size impact the overall task completion time.

It is difficult to compare directly the results of this study with those of the Gould study (1983) discussed earlier because many of the parameters involved differ between the two. Clearly, the type of tasks involved were quite different, with the Gould study requiring participants to compose and edit a letter, and therefore allowed for an open or unrestricted vocabulary whereas the present study used a highly restricted vocabulary. Also, the simulation used in the Gould study did not include manipulation of accuracy rate, and in all conditions simulated a "perfect" recognizer. Finally, one variable of interest in the study was the composition method, where a "draft strategy" allowed users to leave all unrecognized words until a subsequent paper and pencil editing stage, and the "final strategy" required persons to correct all unrecognized words during the letter composition task. Even under the "final strategy" conditions, however, subjects were allowed an editing stage in which they could make any contextual changes. Nonetheless, because these two studies represent the two major efforts to establish the relationship between recognizer vocabulary size and task completion time, a brief comparison is in order. The conditions of the Gould study which most closely resemble those used in this study are the connected speech conditions, using a "final draft" strategy and considering composition time alone, excluding the time spent proof editing. Given these restrictions, only Gould's first study lends itself to comparison, which considered a 1000 word vocabulary (75%), and an unlimited vocabulary (100%). Gould reports approximately 45 percent increase in time in going from a 100 percent available vocabulary to a 75 percent, whereas the results from the present study indicate roughly an 88 percent increase. Despite the differences between the two studies, the results may at first appear alarmingly disparate. At a closer examination, however the difference is at least in part explainable. The current study manipulated accuracy level as well as available vocabulary level. At any given level of accuracy, the *absolute number* of error corrections to be made increases at lower levels of available vocabulary. This point may best be explained by use of an example. If the accuracy level is 90 percent, the available vocabulary is 100 percent, and the task requires 200 words to be entered, only about 20 words (10% of 200) will be misrecognized and require re-entry. If however, the vocabulary is 75 percent, then approximately 50 words (25% of 200) must be entered by spelling them. If the average length of the words to be spelled is 5 characters, then approximately 225 ((5 times 25) plus a "spell" and "end spell" for *each* word) additional inputs must be made. As a result, the total number of utterances is now roughly 425 which means the number of misrecognized words is now 42.5 (10% of 425). In each case, the accuracy level at which the recognizer is performing is 90 percent, however the number of misrecognitions is more than doubled for the smaller vocabulary size. Because the Gould study simulated a 100 percent accuracy level, subjects did not encounter misrecognized words, and in particular did not encounter an increased number of misrecognized words at lower levels of available vocabulary. Understanding this, it should be apparent why smaller available vocabularies appear more favorably in the Gould study than the present study, and that the results of the two studies do not contradict one another.

Although the present study found the available vocabulary level to greatly affect task completion time, it may have still underestimated its significance. Primarily in an effort to minimize subject training time, the orthographic (a, b, c, ...) rather than phonetic (alpha, bravo, charlie, ...) alphabet was used for character level entry. As discussed before, the orthographic alphabet is a *highly* confusable vocabulary for a speech recognizer and would result in very low accuracy rates if a true recognizer rather than a simulation had been used. This, of course, would increase the task completion time due to the time required to correct those errors. Use of the phonetic alphabet, nowever may reduce the speed with which novices can enter character level input and therefore create an even greater difference in time between small and large vocabulary conditions. The Schurick study (1986) confirms that such a difference does exist for novice users of the phonetic alphabet (see also Williges and Dryden, 1987). They report no significance differences between task completion times when the orthographic and the phonetic alphabets were used. Task completion time, however, consisted of time to enter the words as well as time to correct misrecognitions. Because the orthographic alphabet resulted in significantly more misrecognitions, a large proportion of time was spent correcting errors whereas little error correction was necessary when using the phonetic alphabet. The actual time spent *entering* the characters was then greater for the phonetic alphabet than the orthographic alphabet. It is therefore reasonable to suspect that use of the phonetic alphabet in the present study would have further increased the effect of available vocabulary on task completion time.

These results show the importance of using as large a vocabulary size with respect to the task requirements as possible, particularly when lower accuracy levels are expected. This, of course, can be realized in two ways: by using a recognizer with a large vocabulary size or by minimizing the number of words needed for any given application. The latter suggests that a major effort is needed in developing guidelines for dialogue design so that vocabulary size requirements for tasks can be minimized without increasing the processing load on the user. For some applications, however, large vocabularies cannot be avoided, which suggests that research should continue in an effort to increase the allowable vocabulary size of low-cost speech recognition systems.

The relationship between the user's age and task completion time is not as well defined. The main effect of age was not a significant factor, however the Age x Accuracy interaction was. Examination of this interaction indicates that all three age groups performed equally well under the 99 percent accuracy level, but that the younger group performed better under the 91 and 95 percent levels. This is also clearly illustrated in the plots of the response surfaces for each of the three age groups. A possible explanation, and one that is consistent with the studies involving age cited earlier, is that the older subjects entered data as quickly and efficiently as the younger group, however they had a greater difficulty using the error correction procedures. Because the lower accuracy conditions required more error correction, performance for the older group deteriorated more so than for the younger group under these conditions. Examination of the mean task completion times support this idea. Under the "best" condition (accuracy level of 99% and available vocabulary of 100%) in which the fewest number of misrecognitions occurred, the mean task completion time for the older two groups was 7% greater than that of the youngest group.

Under the "worst" condition (accuracy level of 91% and available vocabulary level of 75%) however, the time for the older groups was nearly 17% greater, more than twice that for the "best" conditions. Clearly, the older group was affected more by the increase in number of error corrections required. Also in support of this idea is the significance of the Accuracy x Available vocabulary interaction for groups 2 and 3, but not for group 1. Apparently for groups 2 and 3, lower accuracy levels had a greater impact on task completion time when they were in conjunction with small available vocabularies than with larger vocabularies. As discussed earlier, when the available vocabulary is small, even small decreases in the accuracy level creates a large number of misrecognitions which must then be corrected. The fact that this interaction was not significant for the youngest group indicates that they were less affected by the increased number of misrecognitions.

The fact that the younger group was able to use the error correction procedures more efficiently is likely a result of experience rather than age. The pre-experimental questionnaires indicate that nearly all of the younger subjects report having had either moderate or extensive experience with a text editor or typewriter while groups 2 and 3 as a whole report much less experience. The error correction procedures used in this study are basically a spoken version of those used by most text editors and modern typewriters, and therefore it is not unreasonable to expect that experience with these would transfer to the speech input task. The differences in experience with age is probably exaggerated in this study due to the fact that the youngest group consisted exclusively of university students while the other groups varied more with respect to educational background and occupation. However, because nearly all students are exposed to some degree of computer programming before completing high school, younger people *as a whole* probably do have more experience with text editing and the associated error correction procedures and as a result this trend would likely be found outside this study as well.

Final Error

Accuracy and available vocabulary both had a significant effect on the percentage of uncorrected errors remaining at the end of each trial. A decrease in accuracy level from 99 to either 95 or 91 percent or a decrease in available vocabulary from 100 percent to 87.5 or 87.5 to 75 percent served to significantly increase the number of final errors. Clearly, as the available vocabulary and/or accuracy level decreases, the number of misrecognitions and nonrecognitions increases, therefore increasing the probability that some corrections may be omitted. This effect was partially off-set however, because under high accuracy levels and large available vocabularies, subjects apparently had more confidence in the system and adopted faster entry speeds. In these cases, it was often noted that subjects "entered" a line of information before they detected an error and thus were unable to correct it. Therefore, the difference in error rates between the "good" conditions and the "poor" conditions was not as great as the absolute number of entry errors alone would predict.

It is necessary to note, however that in none of the nine conditions did the average final error rate exceed two percent. And therefore, the "significant" reductions in final error rate due to changes in accuracy or vocabulary level were on the order of 1 percent. It is not likely that in many applications one would choose to gain a 1 percent decrease in final error rate at the expense required to increase the accuracy level from 91 to 99 percent or the vocabulary from 75 to 100 percent. The smaller error rates achieved in this study however, are probably not indicative of all types of tasks involving speech recognition. Subjects generally monitored the visual feedback quite carefully and detected most of the incorrect entries. For tasks which provide less feedback to the user or place demands on the user so they are unable to fully monitor that feedback, the number of uncorrected errors may increase, particularly under conditions of many entry errors (low accuracy or small available vocabularies). Also, regardless of the fact that the instructions stated to work "as quickly and accurately as possible", most subjects took the time to correct all of the errors they detected. In fact, 90 percent of all trials had 3 or fewer errors, which translates to an error rate of 1.34 percent or less. In tasks which are externally-paced or which emphasize quickness as opposed to accuracy, the user may not have sufficient time to correct all entry errors and again, the final error rate will likely vary more with changes in accuracy level and available vocabulary level.

The fact that changes in available vocabulary and accuracy level did not have a major influence on the final error rate is also evidenced by the R-squared value in the regression analysis (R-squared = 0.09). Clearly, accuracy and vocabulary levels alone do not appear to be good predictors of final error rate for the data entry task used here.

Despite the fact that there was generally little variability in the number of errors left uncorrected for a given subject, subjects still varied significantly as to how many errors they left uncorrected. The nonsignificant age main effect indicates this difference was not contributive to the different age groups of subjects studied. Instead, this is probably due to the subjective nature of the instructions. Most subjects apparently interpreted the instructions as meaning to correct all errors, but do so quickly, while a few apparently decided to leave a few misrecognition or nonrecognition errors uncorrected in order to improve their completion time.

Apparently, for tasks which provide adequate feedback and do not overemphasize minimizing task completion time, the levels of accuracy and available vocabulary investigated here affect, but only to a small extent, the number of uncorrected errors remaining at the end of the task. If given the opportunity, most users will correct nearly all of the errors, even under the poorer conditions. This suggests that when minimizing the final error rate is the primary objective, even low quality speech recognizers with lower accuracy levels and smaller vocabularies can provide adequate service. Little is gained in terms of decreased final error rates thru the use of more expensive, higher quality recognizers.

Subjective Ratings

The accuracy level at which the speech recognizer was operating proved to have a significant effect on a user's judgement of whether the device was an acceptable means of data entry for the given task. The youngest group of users identified all three levels of accuracy as being different from one another, with the 99 percent level being most preferred and the 91 percent level being rated as least acceptable. The results from the older groups were consistent in trend, however, the 91 percent and 95 percent levels were not rated significantly different from one another, but both were rated as being less acceptable than the 99 percent accuracy level. Based on these results as well as subjects comments, subjects found having to stop and correct misrecognitions and nonrecognitions to be distracting and quite frustrating. Many subjects also indicated that the particular error correction procedures were not difficult to learn or to use, but that having to interrupt the task to correct errors was disturbing. These results indicate that research efforts are needed to determine the most acceptable form of error correction. Perhaps allowing the user to wait until the end of the task to correct the recognition errors would have caused less user frustration.

Available vocabulary, on the other hand, did not have a major impact on the acceptability rating of the voice recognition system. Again, the trend was the same for both age groups, with the smaller vocabulary levels being rated as less acceptable than larger ones. This effect was only significant for Group 2 and even then post-hoc test failed to locate any difference between the three levels of available vocabulary. From these results, as well as subjects' comments, spelling words character by character apparently was not difficult, nor was it considered to interfere with the task as a whole. Users were easily able to enter spellmode, enter the alphabetic alphabetic characters, and exit spellmode, all without disrupting the "flow" of input.

It is important, however, to recall the exact method of character level entry used. Subjects knew which words were to be spelled before they attempted entry (as indicated by the color coding scheme) and thus were able to avoid trying word-level entry which would then have required deleting the nonrecognition before re-entering the word character by character. Had the subjects not been provided knowledge regarding which words required character level entry, the effects of available vocabulary level on acceptance rating would have likely been greater. As evidenced by the significant effect of accuracy level on user's subjective ratings as well as their comments, subjects found having to stop, back-up, and correct incorrect entries quite distracting, particularly because it required constant monitoring of the visual feedback. In situations where the user must attempt word-level entry before he/she knows to enter the word by characters, a large number of error corrections would be necessary and correspondingly a severe decrease in acceptance could be expected. Also, because the task involved entering printed material, the subjects were not required to know the correct spelling of the words which required character-level entry. In situations which require the subject to spell words from memory, a larger effect of available vocabulary level on ratings of acceptability may be found.

In addition, most tasks would likely employ the phonetic rather than orthographic alphabet for reasons mentioned previously. Because decreased entry rates which are attributable primarily to an increased mental load on the user, are associated with the phonetic alphabet, poorer acceptance by novice users may result as well, particularly for cases which require a large portion of character level entry.

It is interesting to note that the available vocabulary level was not perceived as affecting the acceptability of the system despite its pronounced effect on task completion time. Apparently, subjects were either unaware of the increase in entry time when lower vocabulary levels were used, or they did not consider it significant enough to affect the acceptability of the system. With respect to the second alternative, it is possible that the subjects considered the changes in accuracy level to be so distracting, that in relation, the effects of available vocabulary level were not considered important. This idea is supported by the fact that several subjects expressed extreme frustration with the low accuracy levels, but did not with the small vocabularies.

The importance of the accuracy level and lack of significance of the available vocabulary level is also apparent in the regression analyses. These analyses also indicate, however, that even taken together accuracy and available vocabulary level were not good predictors of the subjective acceptability of the system, according to the low R-squared values of 0.17 and 0.13 (Group 1 and Group 2, respectively). This may suggest that, although significant, neither accuracy nor available vocabulary level is a primary determinant of user acceptance, and that other factors, not controlled in this study, are responsible and should be identified and investigated. On the other hand, accuracy and available vocabulary level may be primary factors, but the rating scale used in this study was not sensitive enough to capture this. Each of the 12 scales actually had a possible range from 1 to 7, which gave the AI measure a range of 12 to 84. Practically speaking, however, because subjects tend not to use the extreme ends of a rating scale (this is referred to as "central tendency error") the effective range for each scale was 2 to 6, which in turn makes the range for the AI 24 to 72. It is possible this range was not wide enough to reflect the true range in acceptance of the different systems used. Because even the subjects' comments suggest that accuracy level in particular had a major influence on their acceptability ratings, future research should concentrate on refining the method of assessing subjects' measures of acceptability and then reevaluating the relative importance of accuracy, as well as available vocabulary on user acceptance.

Age also proved to have a significant effect on the user's acceptance rating. Regardless of accuracy or available vocabulary level, older subjects appeared to rate speech recognition systems more favorably than the younger subjects. This result is consistent with the results of the studies discussed previously, whose authors attribute much of this effect to the fact that older people, as a whole, have less interaction with high technology equipment and therefore have not developed as high expectations. This effect may be slightly exaggerated in this study because of the fact that the two younger groups consisted primarily of undergraduate students, graduate students, and university personnel who probably have an above average exposure to high tech equipment. It was eluded to earlier that there was a practical reason which supports the statistical decision of grouping age groups 1 and 2 together to form Group 1. Because the two groups are similar in their exposure to high tech equipment (at least in this study), which appears to be the underlying cause for the age effect, it also makes practical sense to group them when comparing them with age group 3, which in comparison has had much less experience. It is interesting to note that this effect is present despite the effects of age on task completion time. Older subjects required a longer time to complete the tasks, particularly when the accuracy and/or vocabulary level was poor, however, they still rated the system more favorably than the younger group.

Despite the effects of accuracy and age, even under the worse conditions, the mean rating was favorable (i.e. "neutral" or better on the 7-point bipolar scales). As noted before, the subjective scales may not have been as sensitive as needed to measure the true range of acceptability ratings, nonetheless even the subject's responses were generally favorable. Therefore, at least for novice users, high accuracy systems may be preferable, but less sophisticated systems will be accepted.

Conclusions

The results of this study can be summarized as follows:

- Task completion time is significantly greater for accuracy levels of 91 percent than 95 percent and for levels of 95 percent than 99 percent.

- Task completion time is significantly greater for available vocabulary levels of 75 percent than 87.5 percent and for sizes of 87.5 percent than 100 percent

- Under conditions of low accuracy levels, task completion time is significantly greater for older subjects than younger subjects. No difference in task completion time between age groups exists under conditions of 99 percent accuracy.

- Final error rates are significantly less at higher levels of both accuracy and available vocabulary.

- Subjective ratings of acceptability are significantly greater at higher accuracy levels.

- Subjective ratings of acceptability are not significantly different for the levels of available vocabulary investigated.

- Subjective ratings of acceptabiity are greater for the older two groups than for the youngest, regardless of the condition.

Based on these results, a number of conclusions and recommendations can be made. Because even the small range of accuracy levels investigated here significantly affected task completion time and user's acceptability ratings, the emphasis on the part of system designers and human factors engineers on obtaining high recognition accuracy levels is well warranted and should continue.

The range of available vocabularies investigated in this study had a dramatic effect on task completion time. This indicates that although character-level input may be an effective means of increasing a recognizer's vocabulary, it is not an especially efficient method. Therefore, system designers should continue their efforts to increase the allowable vocabulary sizes of low-cost speech recognition systems. Also, efforts are needed in developing guidelines for dialogue design so that task vocabulary size requirements can be minimized without increasing the processing load on the user. However, until designers are able to develop large vocabulary systems, these results show that character-level entry is acceptable by the user as a means of increasing the vocabulary size, as indicated by the subject's ratings of acceptability.

Older people and/or people less experienced with modern word processing systems had more difficulty with the type of error correction procedures used here, as indicated by the increase in task completion time. Efforts are needed to locate specific areas of difficulty and develop alternate procedures for correcting recognition errors.

Because all subjects reported frustration at having to stop and correct recognition errors, researchers should investigate alternate error correction procedures which will reduce user frustration, for example users may be given the option of correcting errors following a section of text input rather than immediately following each error.

As with any study, some precautions must be adhered to when generalizing the results obtained in this study to other situations. Most notably, these results are specific to the type of task used here and may or may not be applicable to other tasks. For example, the data entry task used here did not have a large cognitive component. The effects of misrecognitions and character level entry on the cognitive process involved in a task such as composing a letter may increase the negative effects of small available vocabularies and lower accuracy rates. As another example, the task used here provided complete feedback and allowed sufficient time for error correction. Under conditions of inadequate feedback or strict time constraints, the effects of available vocabulary and accuracy on final error rates would likely be different. Research is needed to determine how system characteristics such as available vocabulary and accuracy level affect other types of tasks.

In addition, it must be noted that each subject experienced just (2) two hour sessions. If the use of the speech recognition system were to continue over a long period of time, as is the case for many proposed applications, perhaps the acceptability ratings would have indicated an even greater effect due to accuracy and available vocabulary. Therefore, the results obtianed here are applicable to conditions in which the user is a novice or occasional user of such systems, but further research is needed to determine what characteristics are necessary for a system to be considered acceptable by continual users.

In the present study using a simulation of a speech recognition system rather than an actual system proved quite useful, and in fact allowed for the manipulation of variables otherwise not within experimental control. However, when any simulation techniques are involved, some features will likely differ from the actual system. In particular, it should be noted that the procedure of *randomly* selecting words to be "not recognized" (as was the case in the simulation used here) does indeed differ from the manner in which an actual speech recognizer would perform. In any vocabulary, certain words will be more difficult to recognize and therefore more prone to misrecognition. It is quite plausible that users may learn through continued use which words create difficulty for the recognizer and modify their behavior to compensate. For example, if they realize a particular word often results in a misrecognition, they may choose to enter it by characters rather than attempt word-level entry. Because the present study used novice subjets who experienced only a short exposure to the task, it is not likely that they would have adjusted their behavior had the misrecognitions occured in a predictable fashion. However, in studies involving experienced users, or which require the subject to use the system over an extended period of time, the researcher should consider a simulation which chooses it's "misrecognitions" in a probabilistic manner as opposed to random selection.

It should also be noted that the subjects used in this study were all able-bodied. Disabled people, whose use of other input devices is limited, would likely have different criteria for what constitutes an acceptable system. Research is needed to determine what levels of accuracy and available vocabulary are considered acceptable for this population of users. This research would be especially useful since many disabled people have limited incomes and therefore will need to purchase the least expensive system that will still meet their needs.

In conclusion, this research investigated the influence of three characteristics of the human-machine system on three measures of system performance. Although it offers important guidelines both for designers and users of speech recognition systems, much more effort is needed in the area of speech recognition *system* performance. Further research is needed utilizing different types of tasks, other characteristics of the system, as well as other measures of system performance.

References

Baker, J.M. (1982). The performing arts--how to measure up. In D. Pallett (Ed.) *Proceedings of the Workshop on Standardization for Speech I/O Technology* (pp. 27-32). Gaithersburg, MD: National Bureau of Standards.

Bierman, A., Rodman, R. Rubin, D., and Heidlage, F. (1984) Natural language with discrete speech as a mode for human to machine communication. *Communications of the ACM*, 28(6), 628-636.

Chapanis, A. (1975). Interactive human communications. *Scientific American*, 232(3), 36-42.

Clark M.C. (1986). The use of technology in the home by older adults. In *Proceedings of the Human Factors Society 30th Annual Meeting* (pp. 1164-1166). Santa Monica, CA: Human Factors Society.

Cochran, W.G., and Cox, G.M. (1957). *Experimental Designs.* New York: John Wiley and Sons, Inc.

Cohen, A., and Graupe, D. (1983). Speech recognition and control system for the severely disabled. *Journal of Biomedical Engineering* , 2, 99-107.

Cole, A. (1986). Experience with a text editor for spoken input. In *Proceedings of Speech Tech '86,* (pp.149-153) New York, NY: Media Dimensions, Inc.

Coleman, W. D. (1985). Examining the relationship between performance measures and user evaluations in a transfer of training paradigm. Unpublished master's thesis. Virginia Polytechnic Institute and State University.

Coler, C. (1982). Helicopter speech-command systems: Recent noise tests are encouraging. *Speech Technology* 1(3), 76-81.

Connolly, D.W. (1979). Voice data entry in air traffic control (Tech Report FAA-NA-79-20). Atlantic City, NJ: Federal Aviation Administration.

Craft, A. (1982). Human factors and automatic speech recognition: A challenge for all of us. In D. Pallett (Ed.) *Proceedings of the Workshop on Standardization for Speech I/O Technology* (pp. 135-140). Gaithersburg, MD: National Bureau of Standards.

Damper,R.I. (1984). Voice-input aids for the physically disabled. *Int. J. Man-Machine Studies* 21, 541-553.

Davis, K.H., Biddulph, and Balashek, S. (1952). Automatic recognition of spoken digits. *Journal of the Acoustical Society of America* , 24, 637-642.

Doddington, G., and Shalk, T. (1981). Speech recognition: Turning theory to practice. *IEEE Spectrum*, 18, 26-32.

Drennen, J.G. (1980). Voice technology in attack/fighter aircraft. In *Proceedings of a Symposium on Voice-Interactive Systems: Applications and Payoffs* (pp. 201-211). Dallas, TX.

Fink, D.F. (1980). Automatic speech recognition systems, evaluation and application. In *1980 Design Engineering Show and ASME Conference and Seminars.*

Gould, J.D., Conti, J., and Hovanyecz, J. (1983). Composing letters with a simulated listening typewriter. *Communications of the ACM*, 26 (4), 295-308.

Glenn, J.W., Hiller, K.H., and Broman, M.T. (1976). Voice terminal may offer employment to the disabled. *American Journal of Occupational Therapy*, 30, 309-313.

Grady, M.W. (1982). A systems engineering view of advanced speech technology performance standards. In D. Pallett (Ed.), Proceedings of the Workshop on Standardization for Speech I/O Technology (pp.155-158). Gaithersburg, MD: National Bureau of Standards.

Harris, S.D. (1982). Voice-controlled avionics: Programs progress, and prognosis (A case for holistic engineering). In D. Pallett (Ed.) *Proceedings of the Workshop on Standardization for Speech I/O Technology* (pp. 113-129). Gaithersburg, MD: National Bureau of Standards.

Holmgren, J.E. (1983). Toward Bell system applications of automatic speech recognition. *Bell System Technical Journal*, 62 (6), 1865-1880.

Kersteen, Z.A., (1982). An evaluation of automatic speech recognition under three ambient noise levels. In D. Pallett (Ed.) *Proceedings of the Workshop on Standardization for Speech I/O Technology* (pp. 63-68). Gaithersburg, MD: National Bureau of Standards.

Kinkead, R. (1986). Talking to typewriters: Human factors issues and findings in the development of voice-activated word processors. In *Proceedings of Speech Tech '86* (pp. 145-148). New York, NY: Media Dimensions, Inc.

Klatt, D.H. (1980). Prospects for advanced flexible voices and ears for computers. In *Proceedings of a Symposium on Voice-Interactive Systems: Applications and Payoffs* (pp.419-430). Dallas, TX.

Knight, J.A., and Peckham, J.B. (1984). A generic model for the assessment of speech input applications. Final Report, RSRE contract, August, 1984.

Kurzweil, R. (1986). The Kurzweil Voicewriter, A large vocabulary voice activated word processor. In *Proceedings of Speech Tech '86* (pp. 184-187). New York, NY: Media Dimensions, Inc.

Lea, W.A. (1980). Speech recognition: past, present, and future. In *Trends in speech recognition* (pp. 39-98). Englewood Cliffs, NJ: Presntice Hall.

Lea, W.A. (1982). Problems in predicting performances of speech recognizers. In D. Pallett (Ed.), *Proceedings of the Workshop on Standardization for Speech I/O Technology* (pp. 15-24). Gaithersburg,MD: National Bureau of Standards.

Levinson, S.E., Rabiner, L.R., Rosenberg, A.E., and Wilpon, J.G. (1979). Speaker-independent recognition of isolated words using clustering techniques. *IEEE Trans. Acoust., Speech, Signal Processing,* ASSP-27, 336-349.

McCarthy, M.E. (1984). Human factors in voice technology. In F.A. Muckler (Ed.), *Human Factors Review: 1984.* Santa Monica, CA: Human Factors Society.

Meisel, W.S. (1986). Implications of large vocabulary recognition. In *Proceedings of Speech Tech '86* (pp. 189-192). New York, NY: Media Dimensions, Inc.

Morrison, D.J., Green, T.R.G., Shaw, A.C., and Payne, S.J. (1984). Speech-controlled text-editing: effects of input modality and of command structure. *Int. J. Man-Machine Studies* , 21, 49-63.

Moshier, S.L., Osborn, R.R., Baker, J.M., and Baker, J.K. (1980). Dialog Systems: automatic speech recognition capabilities, present and future. In *Proceedings of a Symposium on Voice-Interactive Systems: Applications and Payoffs* (pp. 165-182). Dallas, TX.

Nusbaum, H.C., and Pisoni, D.B. (1986a). Human factors issues for the next generation of speech recognition systems. *Research on Speech Perception*, Progress Report Number 12. Indiana Univ., pp. 405-412.

Nusbaum, H.C., and Pisoni, D.B. (1986b). Automatic measurement of speech recognition performance: A comparison of six speaker-dependent recognition devices. Final Report (IBM Corp. contracts 435114 and 562010), November, 1986.

Nye, M. (1982). Voice integration - the critical mass. In *Proceedings of the Voice Data Entry Systems Applications Conference.* San Mateo, CA: Lockheed Missiles & Space Co.

Ogozalek, V.Z., and Van Praag, J. (1986). Comparison of elderly and younger users on keyboard and voice input computer-based composition tasks. In *CHI '86 Proceedings*, (pp.205-211). Boston, MA: Association for Computing Machinery, Inc.

Osgood, C.E., Suci, G.J., and Tannenbaum, P.H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Pallet, D. (Ed.). (1982). *Proceedings of the Workshop on the Standardization for Speech I/O Technology*. Gaithersburg, MD: National Bureau of Standards.

Pisoni, D.B., Bernacki, R.H., Kubaska, C.A., and Nusbaum, H.C. (1984). Talking in noise: Acoustic correlates of increased vocal effort. *Research on Speech Perception*, Progress Report Number 10, (pp. 171-194), Indiana University.

Poock, G.K. (1980). Experiments with voice input for command and control: using voice input to operate a distributed computer network. (NPS Technical Report, NPS55-80-016). Monteray, CA: Naval Postgraduate School.

Poock, G.K., and Roland, E.F. (1982). Voice recognition accuracy: What is acceptable? (NPS Technical Report, NPS55-82-030). Monterey, CA: Naval Postgraduate School.

Poock, G.K., Schwalm, N.D., Martin B.J., and Roland E.F. (1982). Trying for speaker independence in the use of speaker dependent voice recognition equipment. (NPS Technical Report, NPS55-82-032). Monteray, CA: Naval Postgraduate School.

Rabiner, L.R., Bergh, A., and Wilpon J.G. (1982). An improved training procedure for connected-digit recognition. *Bell Systems Techical Journal*, 61 (6), 981-989.

Rabiner, L.R., and Wilpon, J.G. (1979). Speaker-independent isolated word recognition for a moderate size (54 word) vocabulary. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27 583-587.

Reddy, R., and Zue, V. (1983). Recognizing continuous speech remains an elusive goal. *IEEE Spectrum*, Nov., 84-87.

Rollins, A. (1984). Composite templates for speech recognition for small groups. In *Proceedings of the Human Factors Society 28th Annual Meeting* (Vol. 2) (pp.758-762). Santa Monica, CA: Human Factors Society.

Rollins, A., and Wieson, J. (1983). Speech recognition and noise. In *Proceedings of IEEE ICASSP '83*, Boston, MA: IEEE Press.

Rosson, M.B., and Mellon, N.M. (1985). Behavioral issues in speech-based remote information retrieval. RC11028 (#49528), IBM Watson Research Center, Yorktown Heights, NY.

SAS Institute Inc. (1985). *SAS User's Guide: Statistics, Version 5*, Cary, NC: SAS Institute Inc., 725-734.

Schurick, J. (1986). Efficiency of limited vocabulary speech recognition for data entry tasks. In *Proceedings of the Human Factors Society 30th Anual Meeting (Vol. 2)* (pp.931-935). Santa Monica, CA: Human Factors Society.

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

Simpson, C.A., Coler, C.R., and Huff, E.M. (1982). Human factors of voice I/O for aircraft cockpit controls and displays. In D. Pallett (Ed.), *Proceedings of the Workshop on Standardization for Speech I/O Technology* (pp. 159-166). Gaithersburg, MD: National Bureau of Standards.

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.

Schmandt, C. (1982). Speech communication, a systems approach. In *Proceedings of the Voice Data Entry Systems Applications Conference*. San Mateo, CA: Lockheed Missiles & Space Co.

Spine, T.M., Williges, B.H., and Maynard, J.F. (1984). An economical approach to modeling speech recognition accuracy. *Int. Journal of Man-Machine Studies*, 21, 191-202.

Sweeney, M.J., and Bitar, K.J.(1986). An analysis of friendly input devices for the control of the Naval Warfare Interactive Simulation System (NWISS). Master's thesis. Monteray, CA: Naval Postgraduate School.

Taggart, J.L., and Wolfe, C.D. (1981). Voice recognition as an input modality for the Tacco preflight data insertion task in the P-3C aircraft. Master's thesis. Monteray, CA: Naval Postgraduate School.

Welch, J.R. (1977). Automatic data entry analysis. (RADC TR-77-306). Griffiss AFB, New York: Rome Air Development Center.

Williges, B.H., and Dryden, R.D. (1987). Investigation of high technology devices for the motor-impaired. Virginia Polytechnic Inst. & State Univ., Report Number INF-85-003-01, January, 1987.

Youdin, M., Sell, G.H., Reich, T., Clagnaz, M., Louie, H., and Kolwicz, R. (1980). A voice controlled powered wheelchair and environmental control for the severely disabled. *Medical Progress through Technology*, 7, 139-143.

Zoltan-Ford E. (1984). Reducing variability in natural language interactions with computers. In *Proceedings of the Human Factors Society 28th Annual Meeting (Vol. 2)*, Santa Monica, CA: Human Factors Society.

Appendix A. Personal Information Questionnaire

PERSONAL INFORMATION QUESTIONNAIRE

Please respond to the following questions.

Gender _______Male _______Female

Age _______20-25 _______35-40 _______50-55

Is English your native language? _______Yes _______No

Have you ever used a voice recognition device?

_______Yes, often enough to become familiar with its use

_______Yes, but only in a demonstrational sense (such as at Walt Disney's Epcot Center)

_______Never

How would you describe your knowledge of such devices?

_______Little or none

_______Slight understanding of the principles / applications of use

_______Moderate understanding of the principles / applications of use

_______Thorough understanding of the principles / applications of use

Considering the past year or two, how would you describe your experience with word processors, text editors, or typewriters?

_______Little or no use

_______Small but significant amount of use

_______Moderate amount of use

_______Extensive use

Appendix B. General Description of the Experiment and Participants Informed Consent

Through recent advances in technology, it is now possible to talk to a computer and have it interpret what you say. There are still many problems associated with this technology (that is why you probably haven't seen it used before). This experiment is designed to address a few of the problems. You will be asked to use several different versions of a speech recognition system, nine versions to be exact. Your performance on a task while using each of the nine recognizers will be measured and you will be asked to rate each version on a series of parameters. Note, we are not evaluating you, but rather how performance changes when different versions of the speech recognizer are used. The total experiment will take place on two different days, each day requiring approximately two hours and fifteen minutes of your time for a total of four and a half hours. You will be paid $4.00 per hour.

The experiment will take place in the Human-Computer Laboratory. You will be seated in front of a computer terminal. The task involves reading information from cards and speaking that information into a microphone leading to a computer. The information will appear on the computer's screen for you to monitor. You will correct any errors which may occur. The task may at times seem boring or frustrating. However, no part of this experiment is expected to cause you any pain or harm. Several breaks are scheduled throughout each two and a quarter hour session, at which time you may visit the restroom, water fountain, or simply relax. The experiment will begin again when you are ready. If for any reason you feel uncomfortable and wish to stop the experiment, you may inform the investigator and the experiment will be terminated. You will still be paid for the amount of time that you have spent in the study. However, please note that the research team would appreciate your cooperation in completing the full experiment if you can, so that a full set of data may be obtained.

During the days in which you will be participating, as well as afterwards, please refrain from discussing the experiment with others who may have been or may be subjects in the experiment. Any prior knowledge regarding the experiment may bias a subject's performance, making the results of the entire experiment contaminated. The study is scheduled to end February, 1988, after which you may discuss the study if you wish.

As a participant in this study, you have certain rights. These rights will now be explained to you, and you will be asked for your signature indicating that you consent to participate in this research.

1. As mentioned before, you have the right to discontinue participating in the study at any time, for any reason. If you decide to terminate the experiment, inform the experimenter and he/she will pay you for the portion of time you have participated.

2. You have the right to inspect your data and withdraw it if you choose. In general, data are processed and analyzed after a subject has completed the experiment. At that time, all identification information will be removed and there will be no way to associate your data with you. This is to insure complete anonymity. Therefore, if you wish to withdraw your data for any reason, you must do so immediately after your participation is completed.

3. You have the right to be informed of the overall results of the experiment. If you wish to receive a synopsis of the results, include your address (three months hence) with your signature below. If you should then like further information, you may contact the Human Factors department and a full report will be made available to you.

If you have any concerns regarding this experiment, you may contact either Dr. R. D. Dryden, the principal investigator at 961-6656, or Mr. Charles D. Waring, Chairman of the Institutional Review Board at 961-5283, if you do not feel comfortable talking with the experimenter.

The faculty and graduate student members of the research team sincerely appreciate your participation. If you have any questions about the experiment or your rights as a participant, please feel free to ask. We will do our best to answer them, subject only to the constraint that we do not wish to pre-bias the experimental results.

Your signature below indicates that you have read your rights as a participant and that you consent to participation. If you include your printed name and address below, a summary of the experimental results will be sent to you.

____

Signature

____

____

____

Printed name and address

Appendix C. Detailed Task Instructions

The task you are to perform requires you to enter information into a computer. Please note, however, that **no** previous experience with computers is required. What makes this task unique is that instead of using the computer's keyboard to type information into the computer, you will simply speak into a microphone and a special speech recognition system will interpret what you say and pass the information on to the host computer. Your spoken inputs will immediately appear on a screen positioned in front of you just as if you had typed them, and any commands you enter will immediately be executed.

As you know, when a person types on a typewriter or a computer keyboard, he/she may often make mistakes which must be corrected. Similarly, when you speak to a voice recognition system, it too makes mistakes which must be corrected. You will be instructed as to how to correct these mistakes. Remember, these errors are a natural result of the speech recognition process and do not in any way reflect poor performance on your part.

Task Description
The task itself is similar in nature to one which might be found in the warehouse of a large department store. As merchandise is delivered to the store, it is necessary to make a record of what items arrive so that an accurate record of the store's inventory can be maintained. You will receive a stack of cards representing the shipping labels on arriving merchandise. It will be your job to enter the information contained on the cards into the computer. An example card follows.

item name:	LADIES SILK BLOUSE
ID number:	34628
color:	ROYAL BLUE
size:	8 PETITE
quantity:	6

You will speak each word in order, speaking fairly slowly and carefully pronunciating each. At the end of each line, you will say the word **ENTER**, which instructs the computer to continue to the next line. Each card will contain five lines of information. When you have entered all information on a single card, you will say the word **DONE**, which indicates to the computer that all information pertaining to the current product has been entered, and you will then proceed to the next card.

Your task is complicated by the fact that speech recognition systems are not perfect. Sometimes they confuse one word for another (e.g. you may say "blue" but the speech recognition system thinks you said "two"). In other cases, it may realize you said something, but cannot identify the word at all. In these cases, the system will display on your terminal screen "XXXXXXXXX" indicating to you that it could not understand you. In either of these events, you must back-up and try again. (Exactly how to correct these errors will be explained shortly). It is also possible that the word spoken is not on the list of words that the recognizer knows. Because of the way in which speech recognizers operate, they are only able to recognize a certain number of words which have previously been given to them. Some recognizers can recognize only a few words while others can recognize many. The task you will be performing involves words the recognizer knows as well as words which it does not know. To help you know which words it will understand, these words will be written in black ink on the inventory cards while the words the recognizer does not know will be written in red ink. You will have to use a different procedure in order to enter these words (this procedure will be explained shortly). If you should, by mistake, speak a word which is not one of the words the recognizer knows, it will again display the symbol "XXXXXXXXX" on your terminal screen indicating to you that it did not understand what you said. Notice that when the symbol "XXXXXXXXX" appears on the screen, you cannot always be sure whether the recognizer knows the word but just did not understand you, or whether the recognizer does not know that word.

Correcting Errors

The procedure for correcting these errors will be to first position the cursor over the incorrect item by using the spoken commands **BACK** and **NEXT**. The command **BACK** will move the cursor to the left one word. It may be repeated several times. For example, in the phrase

LADIES SILK BLOUSE

if the cursor is currently positioned over the word BLOUSE, the command **BACK BACK** will move the cursor to the word LADIES. Similarly, the command **NEXT** will move the cursor to the right one word and may be repeated (**NEXT NEXT**) to move the cursor more than one word.

Once the cursor is positioned over the incorrect word, that word may be changed by using one of two methods. You may simply repeat the intended input. If you try this and it does not work (you still get the symbol "XXXXXXXXX") or if you already know the word is not in the vocabulary, you may enter the word by spelling it. First you must say the command **SPELL**, then you may say each letter, being careful to pronunciate it clearly. When the complete word has been spelled, you must say the command **END SPELL**. It is important that you do not forget to say **END SPELL** when you have completed spelling a word.

Correcting Spelling Errors

It is also possible that while you are spelling a word, some of the letters you say will not be recognized by the system. It may confuse two letters (e.g. you say "A" but it thinks you've said "J"), or it may simply not understand which letter you've said in which case it displays the symbol "x" (a lower case letter). These errors are corrected in much the same way as the word-level errors. You will use the command **BACK** to move the cursor over the incorrect letter and simply repeat the letter until the recognizer correctly understands you. While you are in spell mode, the **BACK** command operates slightly differently than when you are entering word-level data. Each time you say **BACK**, the previous letter is deleted. For example, if the cursor were positioned over the "S" in the word LADIES and you said BACK BACK BACK, the letters I E and S would disappear and the remaining phrase would look like this:

LAD

Therefore, you would need to repeat the letters I, E, and S before you said the words **END SPELL** to return to word-level entry. Note that all character-level corrections must be made before you exit the spell mode.

Now remember that each card will contain 5 lines of information. Each line will contain anywhere from 1 to 6 words. While you are entering information on a line, you may make any corrections necessary, but once you say the command **ENTER**, that line of information can no longer be changed. The cursor automatically moves to the next line, ready for more information. You will still be able to see the previous lines, but you cannot change them. So be sure to correct any errors on a line before saying the command **ENTER**.

As you may have guessed, just as the recognizer sometimes misunderstands you as you enter data, it may also misunderstand the command words (e.g. **NEXT, BACK,SPELL,** etc.). The system is designed so that it will not confuse two commands (e.g. **BACK** for **NEXT**), but it will sometimes not understand them, in which case the symbol "XXXXXXXXX" will appear. For example, you may be at the end of a line and wish to continue to the next line, so you say **ENTER**. If, however, the recognizer does not understand the command, instead of returning the cursor to the next line, the line will look like this:

LADIES SILK BLOUSE XXXXXXXXX

To correct this error, you would say **BACK** to position the cursor over the symbol "XXXXXXXXX", and then say **REMOVE** and the X's will disappear. Then, you would say **ENTER** again to continue to the next line. Any of the commands may be misrecognized (**NEXT, BACK, ENTER, DONE, SPELL, END SPELL,** and **REMOVE**) but can all be corrected by removing the unintended X's and repeating the command.

Using these input and editing procedures, it will be your task to enter all of the information on all of the inventory cards into the computer. You will want to work as quickly as possible, but also leave as few uncorrected errors as possible.

The procedures for entering data and correcting errors may at first seem confusing. However, you will be given adequate time to practice using the system, and should "get the hang of it" before the actual experiment begins. The practice session will be similar to the data collection periods, except that the experimenter will be in the room to answer any questions you may have. You should completely understand the task before going any further than the practice session because no questions will be answered during the experiment.

As mentioned before, you will complete 9 different experimental trials over a 2 day period. In each of the 9 trials, your task will remain the same, but the different recognizers may operate slightly differently. In some cases, one may appear not to be able to recognize your voice as well as others. This does not in any way reflect poor performance on your part and you should not try to alter your speech to compensate.

After each trial, you will be asked to complete a questionnaire with regard to the speech recognition system you have just used. This questionnaire along with instructions on how to complete it can be found on the next page. Please review it carefully. On the following page, you will find a table of events, summarizing the experiment and the tasks you will be performing. Review it carefully also. If you have any questions, please ask them at this time. Otherwise, the practice session will begin. Don't forget, you may ask questions during the practice session, but not during the actual data collection trials so make sure you thoroughly understand the task before the end of the practice session. Review the following pages now and inform the experimenter when you are through.

Additional Reminders

- Speak clearly; pronunciate each word distinctly
- Say "ZERO" instead of "OH" for the digit 0
- Say numbers digit by digit ("SEVEN ZERO" instead of "SEVENTY")
- A space will appear between the digits, this is perfectly acceptable (i.e. 7 0 4 6 2)
- Remember not to speak to the experimenter or to yourself during the experiment. The recognizer will attempt to interpret what you say. If you must speak to the experimenter, remove the microphone first.

VITA

SHERRY PERDUE CASALI

HOME:

11 Carriage Hill
Blacksburg, Virginia 24060 (703) 552-6449

Office:

519C Whittemore Hall
VPI&SU
Blacksburg, Virginia 24061 (703) 961-6656

EDUCATION

Master of Science, Industrial Engineering - Human Factors, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.
September 1986-June 1988.
QCA = 4.0/4.0
Thesis: "The Effects of Recognition Accuracy and Vocabulary Size of a Speech Recognition System on Task Performance and User Acceptance."

Bachelor of Science, Industrial Engineering and Operations Research, VPI&SU, Blacksburg, Virginia, 1982-1986.
QCA = major: 3.9/4.0 overall: 3.7/4.0

RESEARCH EXPERIENCE

Summer 1987-Spring 1988: Research Assistant at the Rehabilitation Voice Input/Output Laboratory, Industrial Engineering Department, VPI&SU. Responsibilities included designing, conducting, and analyzing the results of an experiment on the effect of vocabulary size and accuracy level of voice recognition systems on task performance and subjective ratings of acceptability.

1986: Undergraduate Research Project under Dr. Harry L. Snyder, VPI&SU. Designed, conducted, and analyzed the results of an experiment comparing various computer input devices on measures of task performance and user preference.

1985: Senior Design Project under Dr. J. G. Casali, VPI&SU, and J. A. Cregger, I.E. Manager, Hercules Inc., Radford, Virginia. Analyzed current shift rotation schedule of Hercules' employees and suggested appropriate changes. Proposal gained approval from all levels of management.

WORK EXPERIENCE

Spring 1987, Winter 1987: Graduate Teaching Assistant, VPI&SU.

Supervisor: Dean Paul Torgersen, Dean of the College of Engineering.
Course: IEOR 4290, A Theory of Organization.
Responsible for grading quizzes and tests, providing additional assistance to students as needed and occasionally presenting short lectures.

Fall 1986: Graduate Teaching Assistant, VPI&SU.

Supervisor: Dr. Marilyn Jones, Assistant Professor, IEOR.
Course: IEOR 2150, Engineering Economy.
Responsible for teaching three one-hour recitation sessions weekly, grading homework and tests, and providing additional assistance to students as needed.

Summer 1986: Industrial Engineer, Hercules, Inc., Radford Army Ammunition Plant, Radford, Virginia.

Performed a workspace layout for the plant hospital to improve operating efficiency under ordinary, emergency, and disaster conditions. Compared current audiometric testing equipment and procedures with OSHA standards, resulting in the purchase of new audiometric equipment. Conducted a plant-wide lighting survey resulting in the addition or subtraction of flood lights in most areas of the plant. An estimated 5% decrease in electricity consumption was realized.

June 1985 - June 1986: Engineering Aide, Hercules, Inc., Radford Army Ammunition Plant, Radford, Virginia.

Major project focused on the development of a statistical regression model for predicting the overall plant energy consumption. Resulted in the adoption of an energy conservation incentive clause to be added to Hercules' contract with the U.S. Army. Involved extensive computer programming, technical report writing, and formal presentation to top level management.

ACTIVITIES

Honor Organizations

Alpha Pi Mu: Industrial Engineering Honor Society;
Vice President (1985-1986) -- this year the VPI&SU chapter won the National Outstanding Chapter award.

Phi Kappa Phi: National Honor Society

Tau Beta Pi: National Engineering Honor Society

Mortar Board: National Senior Honor Society

Garnet and Gold: Virginia Tech's Junior Women's Honor and Service Organization

Gamma Beta Phi: Honor and Service Organization;
Social Committee (1983-1984)

ACTIVITIES (continued)

Graduate Honor Court: College of Engineering's Representative on the Investigative Panel (1987-1988)

Outstanding Young Women of America

Professional Organizations

Institute of Industrial Engineers;
Recording Secretary (1984-1985)
Graduate Student Representative (1985-1986)

Student Engineer's Council

Human Factors Society

Scholastic Awards and Honors

Cunningham Thesis Summer Fellowship (1987)

Pratt Fellowship (1986-1987)

Pratt Presidential Fellowship (1987-1988)

Graduate Tuition Waiver (Fall 1986-Spring 1988)

John Anderson Memorial Scholarship (1985-1986)

Engineering Senior Scholarship (1985-1986)

Robert P. Davis Outstanding Junior Award Finalist (1985)

Virginia Tech's nominee for the national "IIE Student Award for Excellence" (1986)

Sherry P. Casali

Sherry Perdue Casali

Birthdate: July 19, 1964