A FEASIBILITY STUDY ON THE USE OF A VOICE RECOGNITION SYSTEM FOR TRAINING DELIVERY

by

Marcia Rose Gibson

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

DOCTOR OF EDUCATION
in
Educational Administration

APPROVED:

Bert Wiswell, Co-Chairman

Robert R. Richards, Co-Chairman

Houston Conley

Ronald McKeen

Jimmie C. Fortune

Diane Tierney

April 1990
Blacksburg, Virginia

A FEASIBILITY STUDY

ON THE USE OF A VOICE

RECOGNITION SYSTEM FOR TRAINING DELIVERY

by

Marcia Rose Gibson

Committee Co-Chairman: Bert Wiswell

Committee Co-Chairman: Robert R. Richards

Educational Administration

(ABSTRACT)

This feasibility study examined the possibility of using an independent voice recognition system as the input device during a training delivery requirement. The intent was to determine whether the voice recognition system could be incorporated into a training delivery system designed to train students how to use the Communications Electronics Operating Instructions manual, a tool used for communicating over the radio network during military operations.

This study showed how the voice recognition system worked in an integrated voice based delivery system for the purpose of delivering instruction. An added importance of the study was that the voice system was an independent speech recognition system. At the time this study was conducted, there did not exist a reasonably priced speech recognition system that interfaced with both graphics and authoring

software which allowed any student to speak to the system without training the system to recognize the individual student's voice. This feature increased the usefulness and flexibility of the system.

The methodology for this feasibility study was a development and evaluation model. This required a market analysis, development of the voice system and instructional courseware, testing the system using a sample population from the Armor School at Ft. Knox, Kentucky, and making required alterations. The data collection approach was multifaceted. There were surveys to be completed by each subject: a student profile survey, a pretest, a posttest, and an opinion survey about how well the instruction met expectations. Data was also collected concerning how often the recognition system recognized, did not recognize, or misrecognized the voice of each subject. The information gathered was analyzed to determine how well the voice recognition system performs in a training delivery application.

The findings of this feasibility study indicated that an effective voice based training delivery system could be developed by integrating an IBM clone personal computer with a graphics board and supporting software, signal processing board and supporting software for audio output and input, and instructional authoring software. Training was delivered

successfully since all students completed the course, 85% performed better on the posttest than on the pretest, and the mean gain scores more than satisfied the expected criterion for the training course. The misrecognition factor was 12%. An important finding of this study is that the misrecognition factor did not affect the students' opinion of how well the voice system operated or the students' learning gain.

ACKNOWLEDGEMENTS

There are many individuals who deserve thanks for their assistance during this study. I wish to especially acknowledge the guidance provided by my committee: Bert Wiswell, Co-Chairman; Robert Richards, Co-Chairman; Houston Conley, Robert McKeen; Jimmie Fortune; and Diane Tierney.

I wish to thank John Larson, Bruce Ballentine, Dana Harnish, and Bob Catlett, who were part of the project team, for their support. Special thanks to Gene Winston, Dave Elliott and Terry Tierney.

I wish to thank my husband, Paul, for his loving patience and support; my children for understanding benign neglect; and "Grandma" (Helen Little) for her unswerving devotion to my family.

I also wish to thank my editor, Kathie Parker, for the polish she lent my words.

I dedicate this dissertation in loving memory of my father, George T. Rose (a retired U.S. Army soldier), and to my mother, Elva E. Rose, who have always provided my inspiration. Thanks, Mom and Dad, for teaching me the importance of perseverance, hard work, love of my country and of God.

Table of Contents

List of Figures

Chapter 1

Background of the Problem

Introduction

The use of a voice based computer training delivery system is a relatively new concept for creating a hands-free, computer based training environment (Fuller, 1987). A recent survey of businesses indicated that "fifty-three percent of all organizations with fifty or more employees use computers in their training programs." (Instructional Delivery Systems, 1988, p. 10). Computers are becoming a prime medium for providing training. "There are presently over one million microcomputers installed in the schools across the United States. By the end of the decade this installed base is expected to triple. Speech output has become an integral part of many software programs used in the classroom, especially in the elementary schools and special education." (Adler, 1987, p. 103) However, the use of voice as the input mode for a training situation has not actually gone beyond the research mode. There have been a few attempts in the military to simulate real world requirements by exploring the use of a voice based computer training system. The Air Force, Army, and Navy, to name just a few, have experimented with voice systems for training. The success of the research has not

been overwhelming; and the projects have often terminated without further exploration of the problems discovered during the projects. (Dallman, 1986; Kristiansen, 1985, Bergondy, 1987)

This study was conducted to demonstrate for the U.S. Army's Training and Doctrine command the potential use of voice input for technical skills training. Several applications were considered, such as training armored tank crews in the correct procedures for using the Communication Electronic Operating Instructions (CEOI) to send messages over the radio network. Another application considered for the study was teaching fire commands to the tank crew. These are the oral commands given to the gunner by the tank commander for loading, sighting, and firing the tank's guns. A third application considered for this study was the development of a Russion lesson for the Army's Russian interrogators.

The first two choices were considered because they require the soldiers to speak messages as part of their job performance while their hands and eyes are busy with other important parts of the job. The Russian language is a natural application for a voice system since students need to practice saying the correct pronunciations of a new vocabulary.

For the purpose of this study, the CEOI task was selected. The course currently existed in a computer assisted instructional format using the light pen for input. This would allow for an analysis of the course for changes required to support the voice system but without requiring much change in the content of the course.

The diverse nature of the way humans communicate through speech allows the imagination to visualize applications of a voice recognition system that range from a simple discrimination of yes and no to a speaker-independent, voice-controlled robot. Each of these applications would vary in the required power of the computer, type of software, and interface with the speech recognition pre-processor (Bristow, 1986).

There are many application possibilities for implementing voice interface systems in real word settings (Olson, 1987). For example, the Army is interested in speech technology if the technology can aid soldiers in acquiring the necessary skills for operating military systems and equipment. One group of systems for which a voice based computer training system may be useful is the armor systems. (Chambers, 1987) It is desirable for a soldier to be able to use the voice recognition system from either close range or distance, in

a hands-free environment, with the provision of user mobility. (Wen, 1987) Speech recognition/syntheses applications are essential in eyes-busy, hands-busy environments (Oshika, 1987).

At the time of this study, and even now, a VBTDS is not an off-the-shelf inventory for training material. However, the area of speech technology is rapidly moving and, even as this study is reported, additional progress is being made in the communications software, application areas, and hardware configuration arena. The area of speech technology moves so rapidly that results of this study will be overcome by even more important breakthroughs.

The referenced studies all lent themselves to the basic purpose of this research. This development and evaluation study examined whether an independent voice recognizer could operate in a personal computer as the input for a computer-assisted training situation.

Statement of the Problem

This study was designed to determine the feasibility of using a speaker independent training delivery system to deliver a CAI application designed to train soldiers how to use the Communication-Electronic Operating Instructions (CEOI) code book. This study documented the process for building a voice based computer training delivery system (VBTDS), including design considerations for the development of a training application, integrating it with graphics and courseware authoring software, the operation of the system as an independent recognizer, and the ratings of the trainees of how well the system worked in a training situation.

Research Questions

The study addressed the following questions:

1. What is the current state of the art regarding automatic speech recognition and computer assisted instruction?
2. What are the available options and the minimum hardware and software requirements for a VBTDS?

3. Is a system using a VBTDS as the primary mode of learning a feasibile substitute for current delivery systems used

for training soldiers in CEOI as measured by: a) student learning gains and, b) student ratings of the system.

Significance

The Army was interested in speech technology to assist soldiers in carrying out their missions (Chambers, 1987). The Army funded this study to examine whether a voice based training delivery system could be used to provide instruction. Few studies have been conducted regarding the use of a voice recognition system in a training application. This study was the first of its kind sponsored by the U.S. Army's Training and Doctrine command to explore the use of computer-assisted instruction using independent speech recognition for teaching technical skills. Successful results of this study would allow for the conduct of other studies for other applications suitable for a voice based delivery system such as foreign language training, oral communications, inventory, or record keeping for example. As an added level of significance, this voice system was designed to be speaker independent.

Limitations

This study was limited to examining the requirements for developing a training delivery system using voice as the input and output medium for computer assisted instruction. The courseware taught each subject how to use the CEOI and was developed using current practices of a systems approach to training. The voice based computer training delivery system and courseware was tested using a group of 37 soldiers assigned at Ft. Knox, Kentucky. The soldiers were representative of the target population intended to receive future training using the courseware.

One limitation that affected the study was the scarcity of off-the-shelf independent voice recognition systems. At the beginning of the study no board level recognition system was available and developmental work had to be accomplished to obtain one. This limitation meant there was not a variety of fully developed independent voice recognition systems on the market.

A second limitation was the cost of the new technology. Since this study was investigating new areas in the voice technology arena, associated costs were quite high. The project had a set funding base that could not be exceeded.

A third limitation of the study was the inaccessibility to student performance data for the lecture based CEOI course and the computer assisted CEOI course using the light pen. This data would have allowed for some comparison of student performance across the three types of courses.

The content of the course and the duration of the course provided a fourth limitation of the study since these areas restricted the feasibility of generalizing the results of this study to other types of training.

Yet another limitation of this study was that the population consisted entirely of males. This restricts the generalizing of the results to the female population.

Assumptions

For purposes of the study, it was assumed that the system could be designed and the process of using it to deliver training could be studied. It was further assumed that the subjects used to test the courseware and delivery system did not require familiarity with the operations of a computer. It was assumed that all the subjects had the basic prerequisite knowledge and the ability to read, a familiarity

with military terms, and knowledge of the levels of military echelon.

Definition of Terms

To ensure that the terms used in explaining this study are clear, definitions are provided below.

Automatic Speech Recognizer

A device implementing algorithms for accepting speech as input, determining what is spoken, and providing potentially useful output depending on word(s) recognized (Bristow, 1986). This is the recognition board placed in the computer along with the speech software that interprets the spoken word and places it on the computer screen.

Automatic Speech Recognition

The process or technology which accepts speech as input and determines what is spoken (Bristow, 1986). This is a combination of hardware and software consisting of a voice signal processing board and the software that operates the communications package.

Automatic Speech Recognition System

An implementation of algorithms accepting speech as input and determining what was spoken (Bristow, 1986). This is the technology stored on the signal processing board and within the speech communications software package.

Computer Assisted Instruction

A process of conveying information, practice, and tests by use of a computer.

Connected Words

Words spoken carefully, but with no explicit pauses between them (Bristow, 1986). This would allow the trainee to speak a small string of words into the system within four seconds.

Context Sensitive

Development of a grammar that anticipates available matches based upon the context. The recognizer knows only a limited number of words can follow the recognized word and searches through those templates as opposed to the entire data base.

Continuous Speech

Words spoken fluently and rapidly as in conversational speech (Bristow, 1986). This would be the same as carrying on a conversation with another person. There would be no pauses with no restrictions on the length of the sentences.

Discrete Utterance Recognition

The process of recognizing a word of several words spoken as a single entry (Bristow, 1986).

Enrollment

The process of constructing representations of speech, such as template sets or word models, to be used by a recognizer. This refers to system training as opposed to user training (Bristow, 1986).

Grammar

A scheme for specifying the sentences allowed in the language, indicating the rules for combining words into phrases and clauses. In automatic speech recognition, task grammars specify the active vocabularies and the transition rules that define the sets of valid statements to complete the tasks. The task grammar and structured vocabulary provide syntactic control of the speech recognition process that can enhance performance (Bristow, 1986).

Misrecognition

An example of failure to reject properly spoken input that are not part of the active vocabulary, resulting in selection of a word in the active vocabulary (Bristow, 1986).

Isolated Words

Words spoken with pauses (typically with duration of excess of 200 milliseconds) before and after each word (Bristow, 1986). Each word is said with a small pause such: I-pause-see-pause-the-pause-flower.

Natural Language

Syntactically unconstrained word sequences, typically drawn from a large dictionary and complying with conventional usage (Bristow, 1986). This refers to the words as normally spoken without inserting unnatural pauses or grammar rules.

Nonrecognition

An instance in which a spoken word is ignored, and for which the recognizer or system provides no response (Bristow, 1986). In this case the computer screen remains blank. The system could not find a match in the data base for the word spoken. In this instance the recognition system matched the spoken word incorrectly with the words stored in the data base and put the wrong word on the computer screen.

Speaker-Dependent Recognition

A procedure for speech recognition which depends on enrolling data from the individual speaker who is to use the device (Bristow, 1986). Each speaker must enroll his voice in the data base before the system will recognize his spoken words.

Speaker-Independent Recognition

A procedure for speech recognition which requires no previous enrollment data from the individual speaker who is to use the device (Bristow, 1986). Anyone can speak to the system without adding his voice to the data base.

Training

System training is preferably referred to as enrollment. User training refers to the process of user familiarization

with speech technology (learning how to use an automatic speech recognition device). Training also refers to the conveying of skills and knowledge about the use of the CEOI.

Vocabulary

The words or phrases to be recognized by a recognizer. Distinctions should be made between the complete set of all words or phrases that a recognizer has been trained or programmed to recognize sometimes called the total recognition vocabulary, and the subset of these that may be active at a given time because of an imposed task grammar or other syntactic constraint, called the active vocabulary (Bristow, 1986).

Voice Based Training Delivery System (VBTDS)

An integrated computer system consisting of an IBM-PC clone with 1 megabyte of stored memory, a 40 megabyte hard disk, speech communications board and software, graphics board and software, and an authoring language for writing courseware.

Organization of the Study

This dissertation is divided into five chapters. Chapter one provides a brief introduction to the problem, research

questions, definitions and other preliminaries. Chapter two provides a review of the literature relevant to the overall research objective, and presents a foundation of the research questions listed in chapter one. Chapter three focuses on the method and procedures used. Areas included are discussions of: the population used for the study, the research design, instrumentation, data gathering procedures, and analysis procedures. Chapter four presents the research findings along with accompanying charts and tables. Chapter five presents a summary of the entire research effort, conclusions drawn from the results, and recommendations for further research.

Summary

This dissertation discusses a feasibility study conducted to examine whether a voice recognition system can be used to deliver training effectively. As mentioned earlier, few studies have been conducted on using voice in a training application, therefore, little empirical evidence was available to support the use of a voice based training delivery system. Of the studies conducted, most look at how well the recognizer operates. This study also explored whether a board-level, independent speech recognizer could be

developed since at the time of the study only an external blackbox, independent speech recognizer was available.

Chapter 2

Literature Review

Introduction

This chapter discusses the results of a search conducted by using the key words automatic speech recognition, voice recognition, and speech recognition to locate articles, journals, books, and any other pertinent resources that discussed voice recognition systems and their applications. The search was conducted with the Education Resource Information Center (ERIC) resources, the Defense Technical Information Center (DTIC) resources, and the automated data base system for the Library of Congress.

Background on Automatic Speech Recognition

A review of the literature with regard to the problem of designing a voice based training system disclosed that much research has been accomplished centering around the development of speech recognition systems over the past forty years. (Bristow, 1986; Doddington and Shalk, 1981). Most of the accomplishments have been relegated to isolated, one-word, dependent recognizers. As Woodard and Cupples (1983) indicate, speech recognition is limited to isolated utterance

recognizers which require distinct pauses between utterances and which are speaker dependent. Research efforts seemed to stall in the seventies (Neuberg, 1974) but have enjoyed a resurgence of progress in the eighties. As Peckam states, "The major goals now being addressed by laboratories around the world are: speaker independence, unlimited vocabularies, recognition of fluently spoken words, speech understanding." (Peckham, 1988, p. 387). However, little progress has been made in the area of using an independent voice recognition system as the input mode for a training delivery system. Adler in his introduction remarks at the 1987 Speech Tech conference states:

> There are presently over one million microcomputers installed in the schools across the U.S. By the end of the decade this installed base is expected to triple. Speech output has become an integral part of many software programs used in the classroom, especially in the elementary schools and special education....50,000 Echo speech synthesizers, with the vast majority going to the educational market (Adler, 1987, p. 103).

Adler's comments refer to the speech output mode but do not make reference to the speech input mode. The most success in the speech input/output field has been in the output arena. (Peckham, 1988). A few studies were found that relate to the design of a voice based training system (Levinson and Shipley, 1979; Dickson, Neal and Billingham, 1984; Horn and Scott, 1983). These studies were concerned with the integration of hardware, software, and a voice recognition processor to deliver different applications. The difference between the current study that this dissertation examines and these studies is that this study was concerned with the process for developing an independent voice training delivery system which required the incorporation of a complete system including the independent voice signal processor, the component parts, the performance characteristics, and implications of the integrated system.

Automatic speech recognition is an emerging technology that has not totally evolved from the research and development arena. The technology was first examined in experiments in the Bell Laboratories (Lea, 1980) and has continued to be a technology investigated by the U.S. Government, private industries and international governments. (Martin, 1977), The technology evolved from a black box

technology consisting of the recognizer being external to the computer to the latest technology where the recognizer resides on a board installed inside a personal computer. (Martin, 1977;Lea, 1980, Woodward and Cupples, 1983). The earliest systems were speaker dependent systems which required establishing a voice template for each person speaking to the computer. (Lea, 1980; Doddington and Shalk, 1981). In the 1970s the application of linear predicting coding (LPC) to speech signals was introduced. This analysis appears to be an accurate and efficient method for coding a signal and has enhanced the development of speech technology. (Doddington and Schalk, 1981). The increase in computer power has been suggested as being very important to recent advances in the progress made with automatic speech recognition (Neuburg, 1974; Peckham, 1988). One of the large sources of funding for speech recognition research has been the Defense Advanced Research Projects Agency (DARPA). DARPA has put many millions of dollars into speech recognition research. The ultimate goal of this research was to achieve a speech recognition system that recognizes continuous speech. There are only a few speech researchers currently working in this area. However, IBM seems to have made a significant breakthrough with the Sphinx voice recognizer which was very close to the marketing stage (Neuburg, 1974; Peckham, 1988; Lee, 1988). Isolated word

recognition has been accomplished and is being commercially marketed (Neuberg, 1974; Woodard and Cupples, 1983; Peckham, 1988). Another important research goal of many researchers was to have an independent speech recognizer. There are commercial and government organizations spending millions of dollars to make this possible (Kristiansen, 1985; Bergondy, 1987; Dallman, 1986; Peckham, 1988).

Research in voice recognition has been aimed at the design of speech understanding systems (Klatt, 1974; Levinson and Shipley, 1980). Some research dealing with the application area of voice recognition has been most successful in providing isolated-, dependent-word recognition systems for use in the commercial setting, although governmental agencies were looking at possible military applications (Neuburg, 1974; Nadis, 1988, Fuller, 1988; Woodard and Cupples, 1983, Beek, Neuberg, and Hodge, 1977). Most of the current research did not deal with the use of speech recognition systems in the educational setting (Fuller, 1988). There are a few researchers investigating possible applications for voice recognition such as for teaching foreign languages (Horn and Scott, 1988; Cornick, 1983) and writing skills (Dickson, 1986). Scott Instruments markets a voice based learning system (VBLS) aimed at teaching the physically handicapped the correct pronunciation of words (Horn and

Scott, 1983) which is comprised of an Apple I and a VET series voice entry terminal. This system is designed for use in elementary schools. Cornick (1983) had looked at the VBLS while it was in its development stage as a possible tool for teaching German. She found that the German pronunciation of her test population improved with less than one hour's practice. The research done with regard to the use of a speech recognition system for training was primarily in the military setting. In 1974, the U.S. Navy sponsored research with Logicon to build Flight Training Systems which were also produced for the Air Force. As of 1980, nineteen of these systems were in place and in operation. These training systems used Threshold Technology's VIP-100 speech recognizer. (Woodard and Cupples, 1983; International Resource Development, Inc, 1980, p. 107). The Army has sponsored research for development of training systems with VOTEX for use on Armor equipment. This research did not culminate in systems being fielded. (Kristiansen, 1985).

The types of recognizers on the market today fall into two different categories--speaker dependent and speaker independent. These two types of systems handle a certain size vocabulary, continuous or discrete speech, and can be used for a variety of applications. The major difference in the two systems was that the user of the dependent system

must prerecord the vocabulary to be used with the system before the system will recognize the words spoken. The user of the independent system does not need to prerecord the vocabulary before speaking to the system.

The market for speech recognizers boasts only a handful of manufacturers marketing telephones, wheelchairs, and some inventory systems (Bristow, 1986). The earliest recognizers were described as speaker-dependent systems. These systems required that the voice of the person speaking be recorded and stored on a template. When the person spoke to the computer, the words were compared to the template; and, when a match was found, the action was taken. The systems on the market today are primarily speaker-dependent systems used in industrial settings such as at Lennox China. Lennox China uses a speaker dependent system for inventory and inspection. (Klaver, 1988). Hughes Aircraft uses a system called IMPROVE for data collection in the quality assurance process of manufacturing airborne radar (Ehrens, 1988). Yet another system has been developed by PARTS-DATA to handle voice operated inventory counting (Nelson, 1988). In the case of each of these systems the speaker enrolled the vocabulary by saying each word three or four times. The enrollment was then stored in the database; and when the user said the word during usage of the system, it was matched against the templ-

ate for recognition. These systems usually consisted of small vocabularies which are generally the alphabet and numbers (Peckham, 1988).

The speaker-independent system was the most recent of the areas being researched in regard to voice technology. As Doddington and Shalk (1981) point out, one of three questions often asked about the capabilities of a recognizer is whether or not it is speaker-independent. A speaker-independent system is designed to recognize any speaker without the speaker's voice being previously recorded and stored in the computer. A template that has been created and stored in the computer and words spoken by the user of the system are matched against this template. As Doddington and Shalk (1981) indicated, one factor to consider on the speaker-dependent issue was the sensitivity of the machine to speaker characteristics. They noted that these characteristics tend to experience degradation when speaker-independent recognition is attempted. Peckham (1988) noted that speaker independence was one of the major goals being addressed by laboratories worldwide. He mentioned the SPHINX independent speech system under development at Carnegie Mellon University. This system was based on an acoustic realization of a phoneme that would allow any person to use the system without prior enrollment of his voice. As Lee (1988) noted,

the four important principles on which the SPHINX is based are: (1) a sophisticated yet tractable model of speech, (2) integration of human speech knowledge, (3) utilization of speech units that are trainable, well-understood, and context-insensitive, and (4) capability to learn and adapt to individual speakers. Currently, the first three topics have been researched and implemented. Lee (1988) further noted that the SPHINX system demonstrated the feasibility of a continuous, speaker-independent system consisting of a large vocabulary.

Assessing Performance of Recognizers

A key factor in the voice recognition process is the distribution of energy with frequency (Doddington and Shalk, 1981; Horn and Scott, 1985). The utterance of a sound comes from the mouth cavity and is influenced by the tongue, jaw, and lips. These sounds are referred to as format frequencies. As listeners, we can usually characterize the sound after hearing two or three frequencies. The machine recognizers must do the same thing. Feature extraction, as shown in the figure below, is the first step. It reduces the input to smaller proportions. Linear Predictive Coding (LPC) is used to translate the speech signal into parameters that best fit the speech signal, basically interpreting the sound and making a pattern of it that can be compared to a

reference data base that contains patterns of all the sounds the machine can recognize. Once the patterns are matched, the output is displayed on the computer screen.

So, as seen in the diagram below, the word is spoken (input), frequencies are determined (frequency extraction), the word is digitized and recognition steps accomplished (end-point determination), the input features are made into a spectrum for comparison (pattern formatting) with the spectrum templates stored in the data base (reference data) and a match is made and the word is put on the computer screen (output).

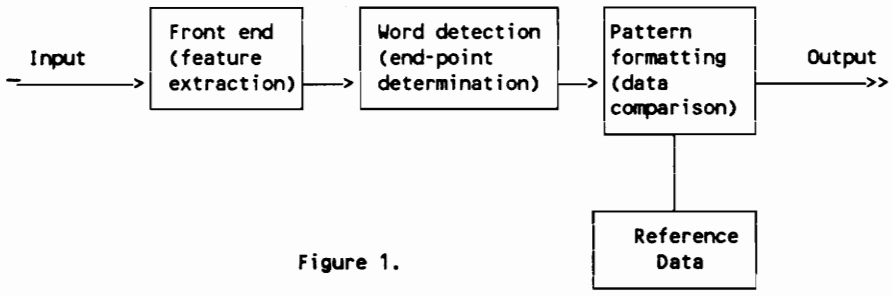

Figure 1.

The performance of the recognizers varied from recognizer to recognizer depending on the conditions under which the recognizer was used. The ideal would be for the recognizer to understand 100 percent of the words spoken to it in a continuous speech pattern. The systems on the market today do not meet this ideal. In order for a system to be considered viable most users are looking for a 95%

recognition rate of discrete speech (Doddington and Shalk, 1981; Martin, 1976). As Klaver (1988) indicated, most of the better systems on the market today give 95-98% accuracy for dependent speech recognizers; but the capabilities of the systems varied.

Factors Which Influence Performance

Factors that affect the performance of individual speech recognizers such as the microphones, the threshold for recognition, the changes in the voice from day to day are evidenced by Emanuelson (1987) who stated, "We found that reliable speech recognition requires that the speaker's mouth stay a fixed distance from the microphone element at all times". The enrollment of the voice for dependent systems may have to be performed more than once. Emanuelson found that the enrollment needed to be updated for different times of the day. A special enrollment of the voice was necessary if the worker came to work with a cold or at some other time when there may be an alteration in the voice pattern. Chambers, Maclay, Gerrits, and Brown (1987) found that it was extremely important to have a robust and reliable microphone. They found that background noise and vibration affected the performance if the microphone was not designed for these conditions. Other factors as pointed out by Promislow,

Larson, Guidry, Eisher, and Joost (1987) were emotional stress, vibration, ambient noise conditions, and the physical condition of the user. Doddington and Shalk (1981) noted that most recognizers commercially available have relatively small vocabularies and frequently confuse words. They also felt that the user needed to learn how to speak to the machine and that the machine's performance varied too widely from speaker to speaker. Martin (1976) separated isolated word recognition systems into two categories - those using high-quality speech and those using low-quality speech. The high-quality speech systems appeared to operate in an environment where the noise level was similar to the laboratory office and the user was performing only the task requiring the voice input. The low-quality speech recognizers were used in factory environments where the user was performing many functions including the one requiring the voice input.

Cornick (1983) discovered that the recognizer did not contain a range of data reference templates to accommodate voices that differed due to age and gender. However, she noted that this was attainable through the voice recognition system. Dickson (1985) noted that each program being integrated into a training system has its own bugs as well as each interface board has its own language. These three

characteristics are inherent factors that affect the performance of the delivery system and the voice recognizer.

Applications for Voice Input for Personal Computers

Some of the applications currently being used for voice recognition systems included automobile assembly line inspection, receiving inspection, automated material handling, parts programming for NC Machine Tools, controls for high performance aircraft cockpit simulators, air traffic control, entry of cartographic and bathometric data, and aids for the handicapped (Bristow, 1986). Speaker identification and verification (O'Shaughnessy, 1986) was another area of research involving voice recognition systems. This process complemented speech recognition in that the attempt was to identify the person speaking rather than what the person was saying. Therefore, "the speech signal must be processed to extract measures of speaker variability instead of segmental features." (O'Shaughnessy, 1986, p. 4) Another application for a voice recognition system was the conversational-mode airline information and reservation system (Levinson and Shipley, 1979). Gottesman (1988) discussed using the voice recognition as a voice mailbox which eliminated the dependency on touch telephones by allowing authorized users to control the system. The user used a password that caused

the system to load the appropriate voice templates so the user could operate the messaging system. Voice systems also help the handicapped as with the Zenith Data Systems' robotic arm commanded by a voice system. It can move paper from the printer. A handicapped person would use voice commands to move the arm. The system performed well to Russian, Spanish, British English, as well as to a Virginian accent (Rash, 1989).

Applications of Voice Input for Training Delivery Systems

The government investigated speech systems for use in the F/A 18 fighter aircraft, for data entry for cartographers (Woodard and Cupples, 1983) and for training situations such as fire commands for the armor tank operator (Kristiansen, 1985). Richard (1982) indicated that the U. S. Navy was investigating using voice technology and artificial intelligence to form an automated instructor's assistant. This idea has been the subject of experiments performed by the U. S. Navy's submarine warfare team. Some issues being researched are the impact on the instructor, loss of personal contact between instructor and student, and appropriateness of message generation for feedback.

Occupational Simulation

Simulation of the operational world in the training delivery system is desired since the processes and procedures can be more realistically simulated (Butler, 1976). The U.S. Army has used a systems approach to training process since 1973 with the introduction of TRADOC PAM 350-30, Instructional Systems Development. This policy-related document set forth the process to develop training that closely emulates the job performance in the work environment. This philosophy continues to be the foundation for the development of training in the U.S. Army and at Headquarters TRADOC as can be seen in TRADOC Regulation 350-7, The Systems Approach to Training (1987). Vocational instruction has required this systematic approach for years and evolved with such notables as Mager, Glaser, Gagné, and Bandura to mention a few (NSPI, 1986; Mager and Beach, 1967). McLagan (1978) describes a structured application activity as one approach to ensure a learning activity transfers to the occupational world along with self behavior modification and a force field analysis technique. Basically, McLagan is recommending training delivery consider motivating information processing and learning in the design and implementation of training. The new skill needs to be used in an application environment that simulates the occupational world (McLagan, 1976; NSPI, 1986; Butler, 1976; HQTRADOC, 1987; Mager, 1988). Studies

conducted by the U.S. Army and other governmental agencies have considered voice input as a means to accomplish some transfer of skills taught to the actual task being addressed (Kristiansen, 1985; Bergondy, 1987; Dallman, 1986).

Summary

This literature review indicated the plethora of research accomplished in the area of speech recognition. As applications become more available on the market, the demand for such systems should increase; and as Martin (1976) indicated, and as has been seen over the last fourteen years, new applications and increased capability of voice input systems have occurred and can be expected.

Chapter 3

Methods and Procedures

Introduction

This chapter outlines the development and evaluation methodology used in this feasibility study concerning the use of voice input/output in the computer aided environment. The following sections discuss: 1) Background, 2) the study population and sample, 3) the study design, 4) instrumentation, 5) data gathering and analysis. A summary section is also included.

Background

The Armor School at Fort Knox, Kentucky trains tank radio operators in proper techniques for communicating over the radio net using the CEOI. The CEOI is a classified code book containing tables listing units and their codes, alphabetical groupings of words and their meanings, radio frequencies, time periods, prowords for sending messages, and other information for sending and receiving messages over the radio net. The CEOI course was taught in the traditional classroom setting using the lecture format. The trainees learned the organization of the CEOI handbook, how to protect

the CEOI handbook, to determine the identification for their unit, to determine the identification of the receiving unit, to encode and decode messages, and to authenticate that the message being received was from a bona fide source, and to enter the radio net. Courseware that teaches this information on a MiccroTiccit delivery system using the light pen for input of course information was used to supplement classroom instruction since the courseware can be used by only one student at a time. When the courseware was used, the trainees generally worked in groups of two. One trainee operated the light pen and the other trainee looked up the information in the CEOI handbook.

Since the Army required that training provided to soldiers be as close to the real world environment as possible, the current lecture approach for teaching the procedures for communicating over the radio net was the lowest acceptable simulation of the real world. There was also light pen courseware which, by itself, did not emulate a real world setting. When the trainees entered the field they looked up information in the CEOI handbook and spoke over a radio. This required the soldier to know the organization of the CEOI and how to look up the codes, which the classroom and light pen courses both handled well. However, the soldier was also required to give the

information over the radio as it was found in the CEOI. The soldier must know the correct pronunciation of the military alpha-numeric phonetic alphabet, to listen before speaking, and to speak clearly and slowly. This was not taught very well in either the classroom lecture format or the computer assisted instruction format using the light pen.

Communication over the radio required the soldier to use the codebook constantly, therefore, a hands-free environment would be required during training the procedures for using the CEOI and communicating over the radio. A voice recognition system could provide a higher level simulation of training to meet this requirement. The trainee could hold the book in his hands, flip through the pages to find the right information, and then speak the communication to the computer just as he would do if communicating over the radio.

Population

The study population investigated consisted of military, noncommissioned officers in the military occupational specialties (MOS) of 19D (Cavalry Scout), 19E (M48 M60 Armor Crewman), and 19K (M1 Armor Crewman). All males, since the Army doesn't recuit women into combat specialties, this population is assigned to tank crew positions and consisted

of four members cross trained to handle all four functions. The radio operator handled the communications for the tank and was required to receive messages, decode messages, encode messages, send messages, authenticate the source messages and enter the communication net. Since any crew member might have to take over this function if a radio operator were critically injured during battle, each member of the crew had to acquire these skills. Therefore, all of the tank crew members are trained to use the CEOI as a part of the Basic NonCommissioned Officer Course at the Armor School. The sample population that was used to validate the use of the voice technology and the course were representative of this population as indicated in the Army's job descriptions listed in Army Regulation 611-201.

Sample

A profile questionnaire was administered to the sample population to obtain information about the group. The questionnaire (Appendix A) was designed to obtain information about the subjects taking the validation course. The questionnaire asked for information concerning the occupation specialty codes of the soldiers, the years in service, the educational background, and prior experience with the CEOI. Racial background was not a question on the student profile form, but observers annotated this information on each

subject's form. The profile questionnaire is typical of one used by the Army for all CAI validations. The results as shown in Figures 2-6 (following pages) are congruent with the description of the target population.

The sample consisted of 37 males in the career management field 19 (Armor) which consists of 19D, 19E and 19K MOS. The sample consisted primarily of E-4 soldiers (48%) although there were representatives from all categories E-3 through E-6 (see figure 2). The mean time in the service was 4.1 years with a standard deviation of 2.6 with about 3.1 years in the military occupational specialty (MOS) with a standard deviation of 2.4. The maximum number of years in Army was 10.5 years; and the minimum number of years in the Army was 1.6 years. The maximum number of years in the MOS was 10.1 years; and the minimum was 0.1 year or not quite two months (See Figure 3) The education level ranged from no high school (5.4%) to some college (21.6%) with the mean education level for high school graduate (67.6%) (see Figure 4). The ethnic composite of the group is shown in Figure 5. The sample consisted primarily of black (24%) and white (59%) males. The rest of the sample was from hispanic (11%), Asian (3%), and American Indian (3%) ethnic backgrounds. One of the subjects had speech problems that he referred to as "lazy tongue" and anticipated having problems with the system. This

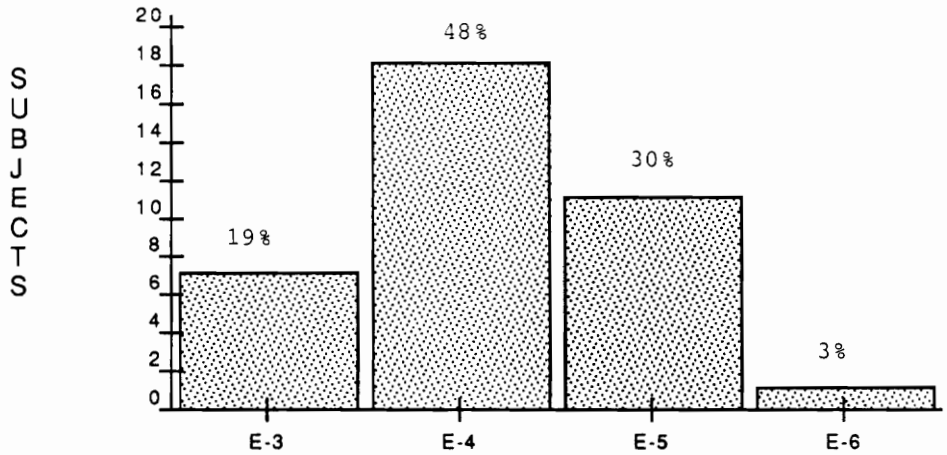

FIGURE 2 - RANK OF SAMPLE

NAME	MEAN	SD	MAX YR	MIN YR
IN ARMY	4.1	2.6	10.5	1.6
IN MOS	3.1	2.4	10.1	0.1

FIGURE 3 --TIME IN THE ARMY AND MOS

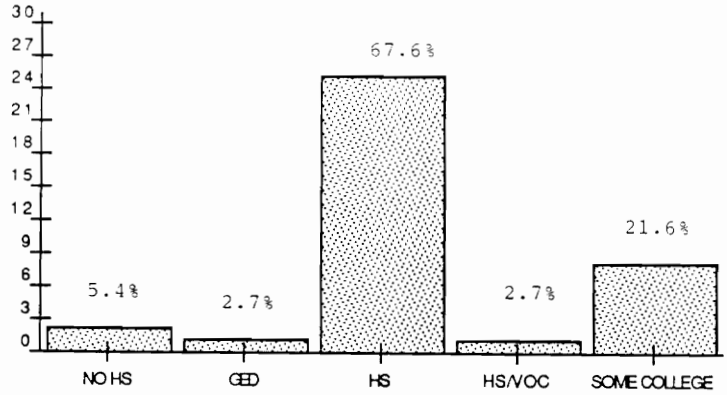

FIGURE 4 - EDUCATION LEVEL

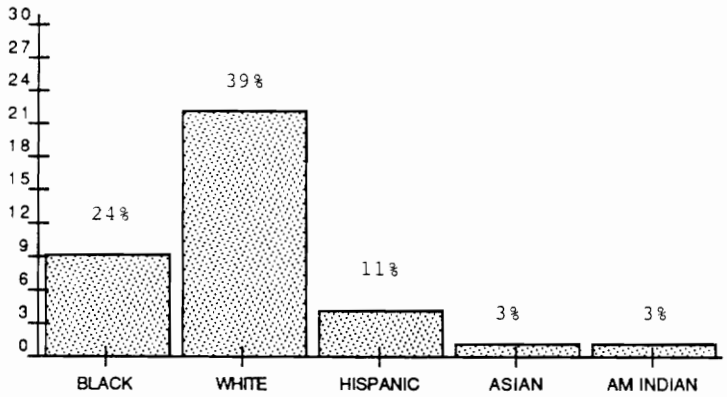

FIGURE 5 - ETHNIC BACKGROUND

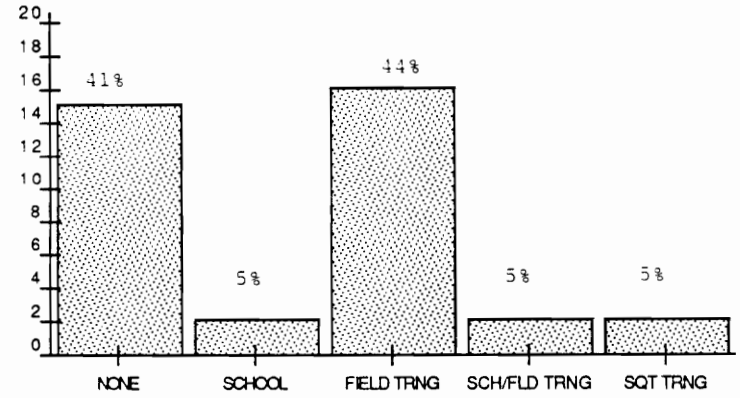

FIGURE 6 - PRIOR KNOWLEDGE OF CEOI

subject was nevertheless used in the study. This question was not asked on the profile questionnaire; however, it was annotated by the researchers on the profile questionnaire as it was submitted by each subject.

Most of the sample had prior knowledge of the CEOI through school training (5%), use in the field (44%), a combination of school training, and field use (5%), or Skill Qualification Test (SQT) training (5%). The remaining 41% had no prior knowledge of the CEOI (see Figure 6).

Procedure

The design utilized for this study was divided into three areas: current state-of-the-art of voice recognition systems, development of the training system and courseware, and evaluation of the training system as a training delivery tool. In order to determine current state-of-the-art in voice technology, a market survey would be conducted. It was determined that the process for developing the training system and the CEOI course would provide the information for development of the voice based training delivery system and courseware. The evaluation of the system would consist of three components: 1) pretest/posttest of the sample, 2) collection of observation data on the performance of the voice recognition system, and 3) opinion survey.

Project Team

A project team was assembled for this study that consisted of a computer specialist from EER Systems, a signal process specialist from Scott Instruments, and an instructional technologist from EER Systems. Support staff for the team consisted of a computer programmer and illustrator from EER Systems. The computer scientist was responsible for integrating the hardware and software components into the training delivery system. His work ensured the integration of the authoring language for the course development with the speech software which would operate the voice output and input. The signal process expert was responsible for developing the board-level, independent speech recognizer and speech communication software. He also trained the other team members on the use of the recognition system. The instructional technologist, the role of this researcher, was responsible for the development of the CEOI computer assisted course using voice input. The instructional technologist reviewed the existing CAI courseware using lightpen and designed a strategy for converting the courseware for using voice. She analyzed the course to determine the vocabulary size and context and supplied this information to the signal process expert. She then designed program flowcharts and storyboards, programmed

the storyboards into the computer using the TenCORE authoring language, developed the evaluation plan, designed the collection instruments, trained the observers for data collection, collated and analyzed the data. The instructional technologist assisted in the collection of voices for the data base templates and co-wrote a paper with the other two primary team members. The team members conducted the evaluation effort, using an additional training specialist from EER Systems to assist in that process.

Market Survey

The concept of the project required a speaker independent recognition system that had a dynamic vocabulary and could handle multiple, selectable vocabularies of fifty words but would allow changes to be made in the vocabulary as training scenarios were developed. Sub-vocabularies and a vocabulary mask to increase recognition accuracy were requirements. The recognition system had to be fully implemented within the context of an off-the-shelf computer based training system. It was also important that an expert CAI developer be able to add voice to lesson development simply as another form of input such as light pen, keyboard or mouse and that all of the features be programmable within an authoring system.

A literature review was conducted to determine what was currently being done in the field of voice technology. The literature review revealed the names of companies, universities, and individuals actively involved in voice technology research as well as key trade shows and symposiums held around the country.

A list was compiled of companies, trade shows, and symposiums to visit. A set of questions was developed to ask each representative about the voice recognition capabilities of each system (See Appendix A). Visits were scheduled to sixteen companies, the 1987 Military Speech Tech Conference, 1987 Speech Tech Conference, and the 1987 AVIOS Conference.

Development of Delivery System and Courseware

The next step in the design was to conduct an analysis of existing courseware for CEOI. The skills to be taught by the courseware were analyzed and course objectives were determined. There was an existing CAI course designed on MiccroTiccit that taught how to use the CEOI. This course used the light pen for input. It was determined to convert this courseware for this project. Thus the analysis of the courseware indicated where voice could be used instead of the light pen as well as how voice could enhance the teaching of the skills for using the CEOI to communicate over a radio.

The vocabulary that would be needed for this project was also determined. Since communications over the radio rely on the use of the phonetic alphanumeric system, the choice of the vocabulary proved relatively simple. The twenty-six phonetic alphabet words and the ten phonetic numeric words were included as part of the vocabulary. The word "decimal" and three other military communication terms were also required for entering answers related to using the CEOI. The other words were control words that allowed the trainee to operate the computer. See figure 7 for a list of the fifty words that were first created for the recognition system.

Once this task was completed, the instructional development was started and flowcharts and storyboards were created. After the storyboards were completed, the graphics were chosen and created using a software package called PC Paintbrush. The course was carefully analyzed for places to use voice for input. The system also had the capability for voice output, therefore, the course was analyzed for places to use the digitized voice output. The source codes were then created. The voice recognition system first used to develop the instruction was a black box version of the board system that Scott was creating. While the courseware was being designed, Scott was building a board level recognizer to fit inside the PC. All the necessary off-the-shelf

VOCABULARY	MODULE 4	MODULE 5	MODULE 6	MODULE 7	MODULE 8	MODULE 9	MODULE 10
CONTROL WORDS							
HELP	X	X	X	X	X	X	X
BACK	X	X	X	X	X	X	X
PROCEED	X	X	X	X	X	X	X
MENU	X	X	X	X	X	X	X
CORRECTION	X	X	X	X	X	X	X
YES	X	X	X	X	X	X	X
NO	X	X	X	X	X	X	X
MIL. PHONETICS							
ALPHABET (26)	X	X	X	X	X	X	X
NUMBERS (0-9)	X	X	X	X	X	X	X
POINT	X	X	X	X	X	X	X
PROWORDS							
THIS IS	X	X	X	X	X	X	X
REFER TO					X	X	X
I AUTHENTICATE					X		X
AUTHENTICATE					X		X
I SET							X
MESSAGE				X	X	X	
GROUPS				X	X	X	
OVER	X	X	X	X	X	X	X
ROGER	X	X	X	X	X	X	X
OUT	X	X	X	X	X	X	X

MODULE 1-ITEM NUMBERS
MODULE 2-CALL SIGNS AND SUFFIXES
MODULE 3-RADIO FREQUENCIES
MODULE 4-CODING AND DECODING
MODULE 5-AUTHENTICATION PROCEDURES
MODULE 6-ENCRYPTING PROCEDURES
MODULE 7- ENTERING THE NET
MODULE 8-POSTTEST

FIGURE 7 - VOCABULARY MATRIX

software was purchased that would allow this training delivery system to meet the concept. TenCORE authoring language was selected for the courseware. Visage Audio software was selected for the recording of voices and digitized output. Catharon software was used to pull all the software together including the voice software created by Scott. Scott provided several versions of the voice software throughout the life of the project. Each version was more dynamic than the last. The template appeared to be more effective in reducing misrecognition and increasing recognition of spoken words.

The CAI developer used the various software packages in the development of the CEOI courseware. There was narration output, feedback output, and exercises and tests that required trainee input using the voice recognition system. The courseware was designed with formative tests after each module and pre and post test mechanisms. Once all of these aspects of the course were programmed, experimental tests were run at Ft. Knox with instructors of the CEOI course. Changes were made to the content of the course to make it more accurate; and recommendations were made about vocabulary words. Several military communication terms were dropped (I Set, Authentication is, and Messages) and one control word (next) was deleted and replaced by four synonyms (forward,

continue, proceed, and go) that would cause the computer to operate in the same manner. Once these changes were made and the board level recognizer installed, the courseware was taken to Ft. Knox for validation of the courseware and verification of the recognizer. The validation was conducted over a three-week period running four students a day through the courseware. Forty subjects were requested for the validation; however, only 37 were available.

After validation the information was collated and analyzed; and changes were made so that the system became more responsive to the training environment. The product was not designed for mass implementation but was put into place in the Armor School for use as appropriate. In keeping with the development and evaluation design, a paper was prepared and presented at the Military Speech Tech Conference in San Francisco and published as part of the Proceedings (Larson, Gibson, Ballentine, 1988).

Evaluation

The evaluation component consisted of obtaining the information required for determining if the speech recognizer could be used in a training situation which required obtaining data about several issues: 1) student performance,

2) observation of recognition accuracy, and 3) student opinion. It was expected that the data from these three situations provided insight about whether or not the students learned, whether or not the students felt they learned, and how well the recognizer performed. Several instruments were used to obtain data for this development and evaluation design.

Pre/Post Tests

Pre and Posttests which the Armor School used for the traditional instruction were used to determine if learning occurred. The Pretest consisted of eight (8) questions (See Appendix A) about the course content the subjects would be learning and was a paper and pencil test administered prior to instruction on the VBTDS. This provided an indication of what knowledge each student already possessed concerning the subject matter. After the course was finished, each student would take a posttest (See Appendix A). This posttest was delivered through the VBTDS in the same manner as the training. It was expected that this would provide information about how well each trainee learned the material just studied. A comparison of the scores on the pre and posttest was intended as an indication that learning did or did not occur for the student.

Observation Form

An observation instrument was also designed for obtaining data on the numbers of recognition, misrecognitions and no recognitions that occurred when the subject spoke answers to the recognizer. The observation form consisted of a data sheet for each module of the course and a data sheet for the control words. Three observers would be used throughout the process. All were trained on how to record the forms prior to the validation. All three observers were given instructions on how to collect the data and the type of interaction allowed with each subject. See Appendix A for the agenda followed for the observer training session. All the words spoken as part of the instruction were recorded on the appropriate module form, and each control word was recorded on the master control form. As each word was spoken, an observer sitting directly behind the student recorded whether the speech recognition system reacted with the accurate word, misrecognized the word, or did not recognize the word at all. (See appendix A for exhibits.) The form was used in a trial situation with four subjects to determine if the format and content would work for the actual data gathering process. No changes were required.

A Cronbach's Alpha was calculated to estimate the internal consistency of the items and revealed a reliability coefficient of .84.

Opinion Survey

An opinion survey was designed to obtain data about how the subjects felt the instruction met the stated training objectives and to obtain data about how well the subjects liked the voice recognition system. It contained thirteen (13) questions about how well the subject felt the course taught what it was designed to teach. A rating scale was used for the answers, and subjects circled the number that matched how they felt about the course. (See appendix A for exhibits.) The form was given to four trial subjects to verify clarity of the questions and the instructions. No changes were required.

Procedures for Data Gathering

The methods and procedures for gathering the data to support this study are discussed below.

Market Survey

As mentioned earlier, the data was gathered for the market survey by visiting representatives of 16 companies, collecting information about the recognizers using the

interview technique. See Appendix A for the interview questions. This involved one trip to the west coast. There were ten companies in the western part of the country. These companies were visited within a 3-week period. The itinerary for the trip allowed basically one to two days at each stop. The other companies were on the east coast and were visited on-site or during the trade shows held in New York City and Washington, D.C. A separate interview sheet was kept for each visit; and all representatives were asked the same questions.

Training Delivery System

The process for developing the system and courseware was annotated in reports which could be reviewed to determine the process.

Evaluation

The data necessary for the evaluation portion of this study was obtained from the instruments described in the above section. For three weeks in August, 1988, the voice system was used to provide instruction to 37 subjects. Forty subjects were requested for the validation period. Forty were assigned for the project but one subject was on leave during the validation period and two other subjects were available for only a portion of the five hours required for

the validation. These subjects were returned to their units. On the first day of the validation, an orientation meeting was held. All of the subjects attended and were given information about the purpose of the validation and the role each would play. At that time each soldier completed the student profile sheet and was given the time for his return to take the instruction.

Orientation

An orientation session was held prior to each student beginning the course. Each subject was told that the microphone was very sensitive and that care must be taken not to blow into the microphone. If the subject wanted to speak to the validation team, he was told to unplug the microphone and then speak. Each subject was told that an observer would be sitting behind him keeping track of how well the voice system performed. The observer would not be able to answer any questions about the course content. Each subject was told to try any word that did not recognize again and that each might need to experiment with how to say the word. Each was told not shout at the machine and to relax and speak normally. Each subject was given a mirror to check the placement of his microphone. Each was told that if the system wasn't recognizing his speech to stop, unplug the microphone, check the microphone position in the mirror, plug

the microphone in again, and try to word again. After this orientation, the subject went to the machine and began the course.

Pretest

The validation schedule was set for two subjects to come in the morning from 7:00 a.m. to 12:00 noon and for two subjects to come in the afternoon from 1:00 p.m. to 6:00 p.m. Upon arrival each student was given a pretest on the instruction to be taught. After the pretest was completed, each subject was given a briefing on how the voice system worked and was told not to move around in the chair or touch the microphone.

Observation

The observer sat behind the subject during the course. He had observation forms on which he recorded whether the words spoken by the subject were recognized, misrecognized, or not recognized at all. (See Appendix A for a sample of the observation form.) If the word was recognized, he put a hash mark on the form under the recognition column. The hash marks were kept in groups of five. If the system misrecognized or did not respond at all (nonrecognition), then the observer put a hash mark under the appropriate column. The observation had an observation for control words

only (words that operated the computer such as forward and back) and one form for each module of instruction. When the subject completed the course, the forms were filed into the subjects data folder that contained all his other data.

Posttest

Upon completion of the instruction, each subject was administered a posttest that asked the same questions as the pretest but using a different set of information. The conditions for the test varied for each subject: each was given a different time period, company to which they were assigned, and company to contact.

Opinion Survey

Each subject was asked to complete the opinion survey after taking the posttest. Each subject was given as much time as was needed to complete the form. The completed form was reviewed with the subject. The subject was thanked for his help and dismissed from the exercise.

Analysis

The analysis for this information consisted of collating the data and summarizing it using descriptive statistics and frequency distributions.

Summary

In summary, a feasibility study was used to investigate the process and procedures for developing a VBTDS. It consisted of a market analysis component which required a literature search and interviews with researchers in voice technology. The process for developing the training system and courseware was documented. The evaluation component consisted of three data collection instruments: Pretest/posttest, opinion survey, and observation collection forms.

The data, analysis, and findings are discussed in Chapter 4.

Chapter 4

Findings

Introduction

This section discusses the results of the Market Survey, pre- and posttest, trainee opinion survey, and observation data for the operation of the recognition system.

Market Survey

The marketing survey revealed that there are voice signal processors being developed and sold commercially; however, most of the effort related to applications for voice output remained in the research and development stages. There were no systems currently being sold that met the criteria for the voice based training delivery system. The desired system required an independent speech recognizer board that could be installed into a personal computer. A survey of current commercial businesses and voice signal engineers disclosed that this board was not currently available.

The Market Survey produced a list of current commercial businesses manufacturing and marketing voice recognizers (See Figure 8). Sixteen companies were found as a result of the

Company	Dependent/ Independent	Discrete Connected	Vocabulary Size	Context Sensitive	Applications	Software Tools	Cost
Kurzweil AI	D	D	1000	Yes	Keyboard, Inventory, Medical, Training	Yes	$6,500
Microphonics Technology	D	D	128	No	Keyboard	Yes	$700
Scott Instruments	D	C*	200	Yes	Training, Inventory QC, Handicapped, Language	Yes	$10,000
Speech Systems, Inc.	D	C	20000	Yes	Natural Language Processing	Yes	
Texas Instruments	D	C	50	No	Keyboard, Telephone	Yes	$1,500
Verbex Voice Systems	D	C	100	Yes		Yes	$6,600
Dragon Systems	D	D	5000	Yes	Keyboard, English, Dictation, Handicapped	Yes	$1,200
Intel	D	D	200	No	QC	Yes	$15,000
Interpath Corporation	D	D	500	Yes	Keyboard	Yes	$600
Interstate Voice Products	D	C	100	Yes	Keyboard, Car Phones, Hospitals, Inventory	Yes	$5,800
ITT Defense Communications	D	C	2000	Yes	Training	Yes	
Keytronic Corporation	D	D	160	Yes	Keyboard	No	$1,000
Voice Connection	D	D	400	Yes	Keyboard, Inventory	Yes	$600
Voice Control Systems	I	D	36	Yes	Security, telephone, Handicapped	No	$100,000
Votan	D	C	300	Yes	Medical, QC, Airline Baggage	Yes	$2,200
Denniston & Denniston	D	C	50	No	Security, Hospital, QC	Yes	

* Scott was in the process of developing an independent voice recognition board and offered to participate in this study.

Figure 8. Market Analysis Results

Market Survey. Each company was asked a series of questions to determine 1) the type of recognizer--dependent or independent, 2) whether the recognizer was a discrete word or connected speech system, 3) the vocabulary size the recognizer could handle, 4) the context sensitivity of the recognizer, 5) applications using the recognizer, 6) software tools and 7) current cost.

Only one company manufactured an independent system. The other 15 companies manufactured dependent voice recognizers. Scott Instruments, however, was in the development stages of manufacturing an independent recognizer. Ten companies were manufacturing discrete systems and six were manufacturing connected speech systems. The vocabulary size of the recognizers ranged from 50 to 20,000 words. Four of the recognizers were not context sensitive. Two of the recognizers did not have software tools. The cost of the recognizers ranged from $600 to $100,000. Three of the companies did not provide cost data.

These systems were reviewed, and representatives of each company were interviewed. The system desired for the project was an independent voice recognizer with discrete speech capability. It was also desirable that the system be able to handle several hundred words, that it could be integrated

into a personal computer along with several other software packages to develop a system, and that it could be used by an instructional technologist to develop courseware without being an expert in signal processing. Most important of all, the system should be affordable. The market survey indicated that the desired system was not currently available. Scott, however, was willing to develop the system for approximately $10,000. Scotts Instruments offered the only system that had been used in language training. Scott's Voice-Based Learning System evaluated student pronunciation and provided corrective feedback. Speaker-dependent with connected speech capability, the system had a vocabulary of about 200 words. It was decided to contract with Scott to build an independent, speaker recognizer at board level. A marketing analysis report was prepared as a part of this project and is included in Appendix B. One note: the technology in speech recognition is advancing so rapidly that a constant update of this type of analysis is required.

Development of Delivery System and Courseware

The information for this component was documented and reviewed to determine hardware and software requirements for the training delivery system as well as the process for developing the courseware. This information is presented in the following sections.

The selection of the hardware and software components was based on several objectives. The hardware had to be compatible with the Army's standards for computer based training delivery systems, the Electronic Information Delivery System (EIDS). All components had to be off-the-shelf and fully supported for updates and maintenance. The software components should be compatible and allow for developing a multimedia training situation using voice input, audio output, graphics, and text.

Hardware

The Army's EIDS system consisted of a PC-AT clone with one megabyte of main memory, a twenty megabyte hard disk, two three and one-half inch floppy disk drives, and a color monitor with supporting EGA and CGA graphics cards. A SONY videodisc player and a Matrox VGA-AT card for overlaying videodisc images with graphics were also part of the system. There were six different input devices with the keyboard and lightpen serving as the primary input devices.

Given these parameters, the hardware components for the Voice Based Training Delivery System follow. The central processing unit (CPU) selected was a PC-AT clone with one megabyte of main memory and a forty megabyte hard disk. Two

floppy disk drives were included, one five and a quarter inch and one three and one half inch. The video display terminal was a multiscan color monitor with supporting EGA graphics card. The inclusion of a Sony Videodisc player and a Matrox VGA-AT card for overlaying videodisc images with graphics and text provided compatibility with the EIDS.

The speech recognition system was a single-board, signal processor with onboard memory for vocabularies of up to 50 words. Any 50-word vocabulary could be downloaded to the board in less than 10 seconds which equated to the time to read a display of textual information. The board accepted speech input from the speaker and converted it into simulated key presses.

The digital audio board used was the Visage 1800E, a single-slot board that created recorded audio messages and stored them on disk. The recordings were used for audio output to enhance the training application process. The Visage board was used as an amplifier that passed spoken words to the Scott signal processor for recognition.

The microphone selected for the system provided a standard headset microphone used with many speech recognizers. It was low cost, lightweight, comfortable to

wear, had noise-cancelling capabilities, and provided stereo audio through earphones that sat just inside the ears.

Software

The software objectives required the training delivery system to be used for authoring of instruction using voice as the input mode. The system was also required to provide the following capabilities: audio output and graphics that were integrated into the personal computer and were compatible with the voice recognition system.

The software selected for the VBTDS included MS DOS 3.2 and its supporting utilities which was the operating system for the computer. The authoring language was TenCORE, a well supported computer-assisted instruction system that supported all of the hardware components mentioned above.

The speech recognition system included software for creating, training and modifying vocabularies as well as two programs that interfaced the speech recognition board with the TenCORE authoring system. These programs were executed automatically when the lesson started.

The digital audio system included two programs that interfaced the audio board with TenCORE. Authoring commands

within the lesson access stored audio messages. The audio board automatically played each message and signal when the message was completed. A special lesson, written in TenCORE by Catharon Productions, orchestrated the recording, playing and timing of the audio. The lesson was specifically designed to interface the Visage board and TenCORE.

The graphics software used was PC Paintbrush, a software package that produced graphic images that were easily incorporated into a TenCORE training scenario.

Process for Developing the CEOI Course

Analysis

The development process followed an instructional systems development approach. An existing CAI lesson that taught the use of the Communication-Electronic Operating Instructions (CEOI) handbook was analyzed to determine how to develop the instruction using the VBTDS. The existing CEOI course used the lightpen as an input device. The analysis of this course determined that the VBTDS would simulate a more realistic world environment for the training situation. The CEOI handbook is intended for use by tank radio operators to transmit messages over the radio network. The CEOI handbook

consisted of unit designators, time periods, and codes that represent certain words. The radio operator needed to know how to use the CEOI to send messages, receive messages, authenticate messages, as well as how to enter the radio network. The lightpen CAI course taught the tank radio operator how to use the CEOI but did not simulate the techniques for receiving or transmitting messages by radio. Since this process required the radio operator to speak, it was determined to be an ideal lesson for the VBTDS.

The review of the CEOI course indicated that the vocabulary needed for this application must contain the twenty-six (26) military phonetic alphabet words, ten (10) phonetic numerical words, five (5) control words for operating the computer, seven (7) prowords (military communication protocol words). These words were provided to Scott Instruments for use in developing the board-level recognizer while the CEOI lessons were being developed and programmed.

Design/Development

The storyboards were written conceptualizing the use of voice as the input mode, the use of graphics to support the storyboard content, and the use of audio output to simulate radio transmissions. The storyboards contained the screen

content to include the graphics and audio output and also indicated the branching flow of the lesson. The branching was linear unless the user was involved in a practice exercise or test and the answer results determined the next screen. The storyboards were reviewed by instructional technologists for the Army at Ft. Knox, Kentucky; and once the storyboards were approved, they were programmed using the TenCORE authoring language.

Since the board-level voice recognition system was being developed at the same time the storyboards and programming were being accomplished, an interim voice based delivery system was used. It consisted of a standalone blackbox system (Scott's VET3) with software enhancements that provided the basic functions of the final delivery system. Because the application interfaces for the interim system and the final board-level system differed significantly, minimizing code changes to the CEOI training scenarios to support the speech functions was important. For this reason a PC-DOS program made speech functions transparent to the TenCORE application by basically emulating the keyboard.

Programming for the use of voice as the input required no extra programming codes. The recognizer returned synonyms for the control words which included special extended ASCII

codes for the function keys. In the TenCORE environment, if the user needed help, he would press the F8 function key and a help screen would appear. Using the recognizer as the input device, when the user said the word "help", the recognizer translated that to the message the computer received if the F8 key had been pushed. This was how the recognizer responded for any of the spoken words. If the speaker were to say the word "alpha" for the letter "A", then the recognizer translated "alpha" to mean pushing the "A" on the keyboard. Thus, when developing the application, thinking of the recognizer as "pressing the keys" on the keyboard allowed the authoring process to incorporate speech into the lesson with minimum coding requirements. The transition from keyboard input to voice input was as easy as flipping a switch.

An audio script was developed, a professional narrator was hired to record the narrations; and the narrations were numbered and coded for integration into the appropriate areas within the application. During the storyboard process, graphics were visualized for the screens. Pictures were selected, given to an illustrator to develop using PC Paintbrush, and then named and coded for integration into the lesson. TenCORE facilitated the actual development and programming of the multivaried aspects of the storyboards.

Template Development

In order for the lesson to work as a speaker independent recognition system, a voice template was required. For the blackbox interim system approximately 25 different voices were recorded and stored. These voices were from Black and White females and males. Developmental tests that were run with this vocabulary indicated a few problem areas. One, the words were presented in a phonetic syntax, i.e. "al-fah" during the training of the vocabulary template. The speaker tended to make the word longer or put the emphasis at the end of the word instead of saying the word naturally. Two, one syllable, short words were also difficult for the blackbox system to recognize, i.e., "go". Three, the vocabulary was trained in alphabetic order and the numbers in numerical order. There was a noticeable difference in the way the speakers said the last words in these sequences, i.e., zulu and niner. Four, isolated training of the words rather than in context seemed to help the recognition of the word. Due to these concerns, some words were dropped from the vocabulary, and a process for training the template was developed.

The blackbox version worked very well; however, when the conversion from the blackbox to the board-level system took

place, unexpected problems were encountered that arose from the system parameter values. That is, during the process of moving from the blackbox interim recognizer,the template would not transfer. As a result of these changes, the vocabulary template had to be recreated just one day prior to validation of the system. Twelve voices were used for this template. The speakers were:

5 females
- 2 White
- 2 Black
- 1 Hispanic

7 males
- 1 East Indian
- 1 American Indian
- 1 Hispanic
- 2 White
- 2 Black

Each of the speakers used to develop the template was recorded three times except for the American Indian; and his voice was used six times. During the training process for creating the template, words were presented in random order, a more natural spelling of the word to be trained was used, and each speaker was told to place each word into a military communication context.

In summary the development process for a lesson using the VBDTS would consist of the following steps:

- analyze the training situation
- determine if the training situation required voice input
- determine vocabulary words, sub-vocabularies, and synonyms
- determine masking requirements
- determine audio output requirements
- determine graphic requirements
- develop storyboards
- program source codes integrating audio output, graphics and text
- train the vocabulary to the system

Evaluation

This section provides the findings of the three instruments: 1) pretest/posttest, 2) opinion survey, and 3) observation data on the recognition performance of the VBTDS.

The VBTDS performed very well during the validation process at Ft. Knox, Kentucky. The objectives were 1) to deliver instruction to the subjects without anyone failing to complete the lesson because of the system, and 2) to obtain no more than a 15% misrecognition factor. The findings are described below.

Pretest/Posttest Results

Each subject was administered a pencil and paper pretest and a VBTDS posttest on the instruction with each test consisting of eight questions. The results on the pre and posttest reflect the performance of thirty-four of the subjects. The pre and posttests for the remaining three subjects were not included because of missing information or improperly completed forms. Figure 9 shows the mean pretest/posttest results as well as the mean gain. The best score they could have made was an 8. The subjects had a mean score of 2.97 with a standard deviation of 1.96 on the pretest. The subjests had a mean score of 6.15 on the posttest with a standard deviation of 1.52. A comparison of the two indicated a mean gain from the pretest to the posttest of 3.18 points with a standard deviation of 1.84. As can be seen, the group performed better on the posttest than on the pretest. Figure 10 shows that 85% of the subjects increased their scores over their pretest scores. The other 15% had no change in the score they made on the pretest and the posttest. No one did worse on the posttest than on the pretest.

Observation Data Results

Data was gathered about whether the word spoken was recognized, misrecognized, or not recognized. A total of

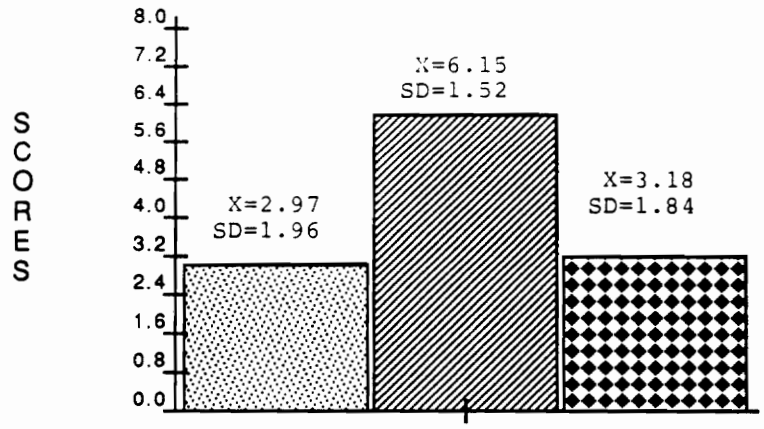

FIGURE 9 - PRETEST/POSTTEST/GAIN RESULTS

PRETEST POSTTEST GAIN

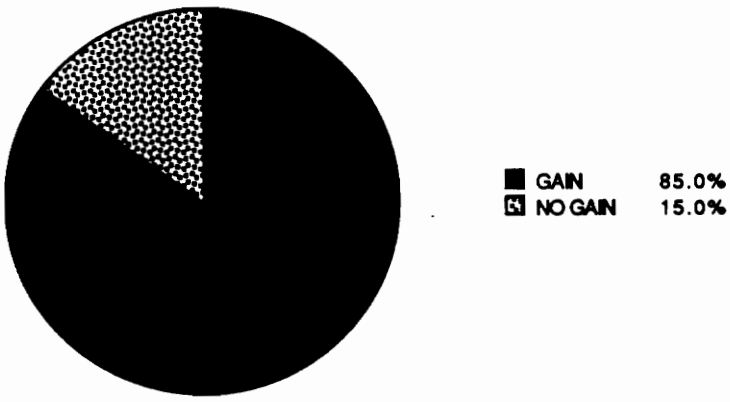

FIGURE 10 - PRETEST/POST TEST RESULTS

29,871 words were spoken by the thirty-four subjects. The data for only 34 of the subjects was used because of incomplete information on three of the subjects. Also, the total of words spoken by each individual differed depending on how many words misrecognized or did not recognize. If a subject had to repeat a word because of no recognition, then he said an additional word. If the system misrecognized, the subject had to say "correction" for each word misrecognized and then say the correct word. These factors caused the total number of words spoken for each individual to be different. As depicted in Figure 11, the mean misrecognition level was 195 words, with a standard deviation of 138 and a range of 435. The percent shown indicated a mean of 12% The results indicated that 72% of the words spoken were recognized, sixteen percent of the words were not recognized, and twelve percent of the words were misrecognized (See Figure 11). The sixteen percent of the words not recognized was not considered a problem. This meant the system did

N=34	MEAN	SD	RANGE	MEAN PERCENT	PERCENT RANGE
RECOGNITION	802	136	1023	72	36
NONRECOGNITION	139	82	757	16	29
MISRECOGNITION	195	138	435	12	20

Figure 11. -- Observation Data for Recognition System Performance.

nothing when the word was spoken, indicating the system did not "hear" the word. The important descriptive statistic here was the 12% misrecognized which indicated that the system did not recognize the word and gave the student the wrong response so that the student needed to erase the wrong word from the screen and then say the desired word once again. The desire at the outset of the study was to develop a system that performed with no more than a 15% misrecognition level. This goal was not only met but exceeded.

Opinion Survey Results

Each subject was administered an opinion survey to determine if they thought the instruction taught the objectives of the course and to determine what they thought of the training delivery system. A Cronbach's Alpha was calculated to estimate the internal consistency of the items and revealed a reliability coefficient of .84.

The Trainee Opinion Survey indicated that, overall, everyone felt the instruction and delivery system met the course objectives very well. Thirteen questions were asked. The subjects were asked to respond using a Likert-scale, with 1 representing the low end and 5 representing the high end of

the scale. The last question on the survey asked for a slightly different ranking. The subjects were asked to rate the difficulty of the course as 5-4 (too high), 3 (about right), and 2-1 (too low).

Figure 12 depicts the mean score for each question with the last bar showing the overall mean rating which was 4 (an overall rating of very good). Figure 13 illustrates the mean and standard deviation for each question. As can be seen, the mean scores all rated the course and the system as performing very well. The mean score of 2.97 for question 11 indicated that the subjects had very little prior knowledge of the CEOI. The mean score of 3.06 for question 13 indicated that the subjects thought the difficulty of the course was about right.

The results of this opinion survey were very positive in terms of how the subjects felt about the instruction and the delivery system. The opinion of the subjects indicated that the system can be used to deliver training.

Also, "Question 11: How would you rate your knowledge of this material prior to taking this lesson?" had a mean response of 2.97 indicating that, on the average, the subjects thought their knowledge of the CEOI handbook was

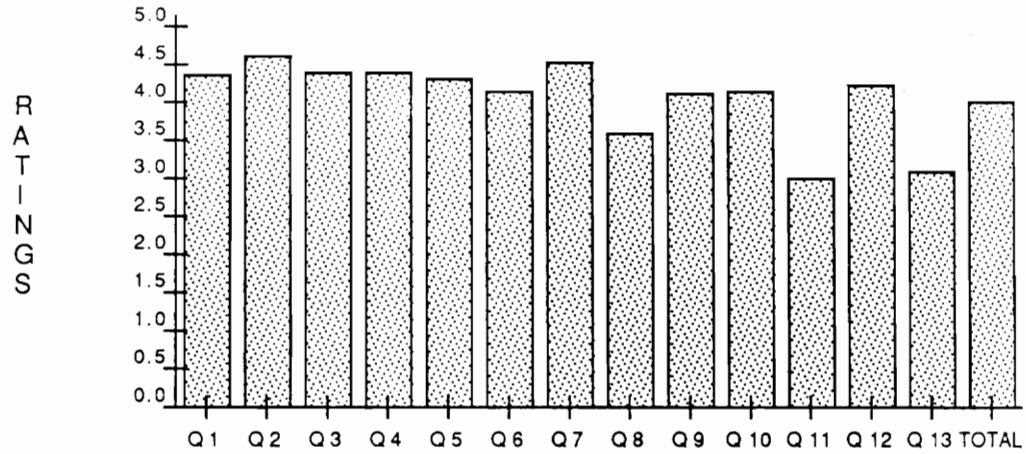

FIGURE 12 - OVERALL OPINION SURVEY RESULT

QUESTION	MEAN	SD
1-MET EXPECTATIONS	4.32	.67
2-INSTRUCTIONAL ORG.	4.60	.51
3-CLEAR OBJECTIVES	4.37	.71
4-INSTR/OBJ MATCH	4.37	.79
5-SAMPLES/PRACTICES HELP	4.28	.69
6-PICTURES HELP TO LEARN	4.12	.74
7-INSTR/TEST MATCH	4.49	.60
8-OPERATION OF VOICE SYSTEM	3.56	.84
9-ORIENTATION TO COMPUTER	4.09	.87
10-RATE VOICE SYS FOR TRAINING	4.11	.87
11-PRIOR KNOWLEDGE OF CEOI	2.97	.94
12-CEOI KNOWLEDGE AFTER COURSE	4.19	.62
13-DIFFICULTY OF COURSE	3.06	.41

FIGURE 13
OPINION SURVEY RESULTS

fair. A closer look at the responses to this question indicated that 21 of the subjects rated their previous knowledge of this material as being fair to having none at all. Of these same twenty-one (21) subjects, ten (10) scored three (3) or less correct on the pretest and scored four (4) to eight (8) correct on the posttest. This lends some credibility to the subjects' general opinion that their knowledge before the instruction was fair.

A closer look at "Question 12: How would you rate your knowledge of this material after taking this lesson?" indicated that a mean response of 4.19 thought they knew more about the material after the course. Twenty-eight (28) of the subjects thought they knew more about the material after the course. Twenty-five of these same subjects increased the scores they made on the pretest when they took the posttest. The other three subjects scored the same on the pretest as they did on the posttest but still thought they knew more about the material after completing the course. These statistics, even though descriptive, lent some credence to the subjects' general opinion that their knowledge of the material increased.

A closer evaluation of "Question 8: How well did the voice recognition system operate for you?" revealed a mean

response of 3.56 (between fair and very well) with a standard deviation of .84. Fifteen subjects rated the operation of the system with a 3 (fair) and another 13 with a 4 (very well). Four subjects rated the operation of the system with a 5 (excellent). Only two subjects rated the operation of the recognizer low -- one with a 2 (poor) and one with a 1 (not very well).

Subjects 3, 12, 24, and 35 rated the operation of the recognition system as excellent (5). The recognition data shown for subject 3 indicates a low misrecognition percentage for all four of 4%, 11%, 5%, and 12% respectively. The two subjects who rated the system low had high misrecognition scores of 23% each. However, subject 20 had an even higher misrecognition level of 24% and rated the system as fair.

An important fact here is that the misrecognition factor did not appear to prevent either subject 9 or 11 from gaining knowledge from the course since they both showed a gain between the pretest and posttest of 4 and 3 points respectively. An examination of the data on subject 20 indicated a gain between the tests of 2 points. This appears to support the precept that the recognition system can be used as an input means in a computer assisted instructional environment.

Figure 14 depicts the results of Pearson Product Moment correlation on the relationship between Question 8 and the misrecognition factor, Question 10 and the misrecognition factor, and the gain in knowledge and the misrecognition factor.

There was a correlation of -.21 between Question 8,"How well did the voice recognition system operate for you," and the mean misrecognition percentage as shown in Figure 14. This would indicate a small but not significant relationship between the students opinion about the operation of the voice system and misrecognition factor. The inverse relationship would be expected, however, it would not be expected for the correlation to be so low. This would appear to indicate that the misrecognition factor had very little impact on how the students felt the voice system operated.

There was a correlation of -.03 between Question 10,"How would you rate the voice system as a training application for this type of lesson," and the mean misrecognition percentage as shown in Figure 14. This would indicate no systematic relationship exists between the students' opinion that the voice system could be used as a training application and the misrecognition factor. This finding is very important since

COMPONENTS OF CORRELATION	PEARSON PRODUCT MOMENT (N=33)
MISRECOGNITION/OPERATION OF SYSTEM	-.21
MISRECOGNITION/USE SYSTEM TO TRAIN	-.03
MISRECOGNITION/GAIN IN KNOWLEDGE	+.28
GAIN IN KNOWLEDGE/OPERATION OF SYSTEM	+.13
GAIN IN KNOWLEDGE/USE SYSTEM TO TRAIN	+.17

FIGURE 14 -- CORRELATION:
SYSTEM OPERATION / USE TO TRAIN / GAIN IN KNOWLEDGE TO MISRECOGNITION PERCENTAGE

it might be expected that the misrecognition factor would cause the students to say the system should not be used for training; however, just the opposite occurred. Another important aspect about this finding would be that the misrecognition factor does not have to be extremely low for the students to accept the voice system as an application for training.

There was a correlation of +.28 between the students' mean gain of knowledge score and the mean misrecognition percentage as shown in Figure 14. This would appear to indicate a small but not significant relationship between how well the students performed on the posttest and the misrecognition factor. Once again, a direct relationship would be expected between students' gain in knowledge and the misrecognition factor. This correlation data seemed to imply that this relationship is not very important which further supported the idea that the voice system could be used to deliver training.

There was a correlation of +.13 between Question 8,"How well did the voice recognition system operate for you," and the students' mean gain in knowledge as shown in Figure 14. This would indicate a small but not significant relationship between the students' opinion about the operation of the

voice system and the gain in knowledge. This would appear to indicate that the students' gain in knowledge factor had very little impact on how the students felt the voice system operated.

There was a correlation of +.17 between Question 10, "How would you rate the voice system as a training application for this type of lesson," and the students' mean gain in knowledge as shown in Figure 14. This would indicate a small but not significant relationship between the students' opinion that the system should be used for training and the mean gain in knowledge. This would appear to indicate that the students gain in knowledge had very little impact on their opinion that the VBTDS should be used for training.

These findings were considered important because they appeared to support the premise of this study that a voice recognition system could be used for delivering training. Furthermore, it appeared that the misrecognition factor did not affect the students' opinion on how well the system operated or whether it should be used as a training application.

In summary, the general assessment of the Trainee Opinion Survey was that the subjects thought learning occurred which supports research question 3 and that the recognition system operated well which partially answers research question 4.

Enhancing and Limiting Features of a Voice Based Training Delivery System

This prototype voice-based training delivery system provided a new dimension to CAI and allowed the training to simulate real world job performance criteria more closely than ever before. Some enhancing features of the system are that the system contains a board-level independent voice recognizer, the system integrated the voice recognizer graphics software, audio output, and had an instructional authoring capability for the purpose of delivering training, and it was relatively easy to use.

The development of the CEOI lesson took only eight weeks and included creating and programming the storyboards. The storyboards required the integration of graphics, audio output, and voice input.

Creating voice templates was a relatively simple task. On the average it took approximately thirty minutes to program a voice into the template for a fifty-word vocabulary. The process required each person to say each vocabulary word three times so that each person's voice became part of the template. The template could be created with as few as twelve voices. Although memory would be a limiting factor, more voices could be added and would enhance the database of the recognizer. The system could then recognize any speaker and thus be called an independent recognizer.

The cost of the entire system including the AT-compatible CPU, 40 megabyte hard disk, EGA color monitor, board level recognition system, authoring system, and audio/recognition libraries was $7500. The voice recognition system used a full slot board with 40 megahertz processor that was speaker independent. It had 160 active words or phrases with the ability to swap vocabularies in seconds, 32 active vocabulary masks that allowed the instructional designer to mask out all words but a few in certain parts of the lesson. It contained a high fidelity digitized audio capability of 32,000 samples per second. Furthermore, the recognizer had a graphic audio editing capability for cut and paste of the template.

The development library allowed for speech characteristics to be dynamically controlled. A silence threshold was set which allowed for acceptance of certain background noises. An acceptance threshold and an acceptance delta was set for each word. The system allowed for the programming of embedded silence as well as for beginning and ending silence for a word. Two other features of the system were application synonyms and adaptive recognition. The use of the recognizer in developing the independent voice template was very simple. These features are significant since they affect the accuracy of the recognizer's acceptance of the words spoken.

Another enhancing feature of the system was the ease of use for subjects. Each subject took approximately 10 to 15 minutes to become comfortable learning to communicate on the system. The subjects were able to work with the course materials and the computer simultaneously. They could look up information in the codebook and then say it without having to copy the information down and enter in with a keyboard or lightpen. This process more closely simulated real world job requirements.

Another feature of the system was the ability to create computer generated graphics using PC paintbrush software. The graphics were relatively easy to create and could be input to imitate motion.

The use of audio output helped to create a more realistic training scenario, especially for this training situation. Voice output was used for narration, feedback, and to simulate receiving radio messages.

One limiting feature of the system was the headphone which was extremely sensitive to movement and touch and required the subjects to remain as motionless as possible in their seats. The subjects had to be aware that they could not touch the microphone or adjust the headset while the voice recognizer was on. Another limiting feature of the system was the lack of an on/off switch that allowed the voice system to be turned off when not needed. If the subjects wanted to stop during the lesson, they unplugged the microphone to stop the system.

SUMMARY

To summarize, the voice based training delivery system operated acceptably in a training situation. All of the subjects completed the lesson material with an average misrecognition accuracy rate of 12%, which was better than the objective of the research. However, the importance of nonrecognized words was not considered at the onset of the research study. The nonrecognition factor was 16%. This factor did not cause the subjects any real problems since they merely said the word again. This factor was important since the subjects completed the course. Since nonrecognition means the system did not hear the word and the subject must repeat the word, this factor is added to the recognition accuracy. The system did not misrecognize the subject and cause him to have to erase an error or miss a question.

The subjects rated the system very well in the delivery of a training application. The subjects rated the system between fair and very well for recognition accuracy which further indicated that the recognition factor did not discourage the subjects about the capability of using the system for delivery of training.

Chapter 5

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

A summary review of the research conducted was provided and conclusions drawn from the study findings are presented. Recommendations are offered for implementing the study conclusions and to aid future research.

Summary

This feasibility study was concerned with determining the current state-of-the-art of the voice recognition system industry and whether or not an independent voice recognition training delivery system could be developed to present training. In addition, the study was designed to determine what the instructional development process would resemble within this medium. A market survey was conducted and indicated that there were only a small number of commercial vendors selling independent voice recognition systems and that few training applications had been successfully developed. Thus, the research reported in this dissertation represented a step into that dimension.

The central question asked in this study was: "Can an independent voice-based training delivery system be developed for use in presenting training?" Prior to this study, no successful training applications could be documented that used an independent voice recognizer in a PC environment.

Background

The U. S. Army was searching for an effective medium for training military commanders on the use of the Communications Electronic Operating Instructions (CEOI) Codebook to communicate in a radio network during military operations.

The information in the CEOI allows the commanders to encode data to be transmitted, to decode data received, and to locate and identify units for whom the communications are intended.

Currently, this skill is trained using a traditional, platform-lecture format. A computer-based training courseware using the lightpen for input is used to supplement that training. Maximum training effectiveness is achieved when the instruction approaches the actual performance required of the tank commander. For this situation, a hands-

free environment is required as well as a requirement to speak over a radio network in code.

Methodology

A feasibility study using the development and evaluation approach was used to conduct this study. It consisted of three components: 1) a market survey, 2) developoment of a training delivery system, and 3) evaluation of the training delivery system and courseware.

Market Survey

An extensive market survey was conducted across the country to determine the current state-of-the-art in independent voice input technology. Sixteen companies were visited to obtain information about their voice recognition systems. No off-the-shelf alternatives were discovered; and the development of an independent voice recognizer was undertaken as part of this study.

Scott Instruments was selected as the company to develop the independent voice recognizer. They currently had a "black box" independent recognizer called VET3. They were in the process of developing a board-level recognizer to insert

into a personal computer. The requirements for the delivery system were determined and integrated into a voice-based delivery system. The system was field tested at Fort Knox, Kentucky with 37 subjects. All of the subjects completed the four hours of courseware and gave the delivery system a high rating as a mechanism for presenting training. One note: the technology in speech recognition is advancing so rapidly that a constant update of this type of analysis is required.

Training Delivery System

The training delivery system should allow the trainee to hear audio examples of voice communication, to examine visual displays of graphics, text and combinations of the two, and to enter all responses using voice communication. The system must contain an authoring language which would allow the development of courseware integrating graphics and audio input/output.

Hardware

Because of the requirements stated above, a system was built which consisted of a PC-AT clone central processing unit with one megabyte of main memory and a 40-megabyte hard disk drive. Two floppy disks were included (one five and a quarter inch and one three and one-half inch). The video display terminal was a multiscan color monitor with

supporting EGA graphics card. A single-board housed the independent voice recognition system which consisted of a signal processor with onboard memory for vocabularies of up to 50 words. Any 50-word vocabulary could be downloaded to the board in less than 10 seconds. The digital audio board was the Visage 1800E, a single-slot board that created recorded messages and stores them to disk. The board also provided for the recording andn playback of any speaker's voice. The Visage board was used as an amplifier that passed spoken words to the Scott signal processor for recognition. The microphone was a standard headset, lightweight, comfortable to wear with voice-cancelling capabilities.

Software

The software integrated into the training delivery system consisted of MS-DOS 3.2 and its supporting utilities, TenCORE authoring language, and Scott's speech recognition software which included software for creating, training, and modifying vocabularies. It also included two programs that interfaced the speech recognition board with the TenCORE authoring system. The digital audio system included two programs that interfaced the audio board with TenCORE. A special lesson, written in TenCORE by Cotharon Productions orchestrates the recording, playing, and timing of the audio. This lesson was specifically designed to interface the Visage

board and TenCORE. The graphics system used was Paintbrush, a system that created graphics easily incorporated into the TenCORE training scenario.

Courseware

The development process followed an instructional systems development (ISD) approach that was used for an existing CAI lesson that teaches use of the CEOI. The lesson, programmed in TICCIT, used lightpen for input. The courseware was analyzed to determine the required vocabulary and where voice would be used. All existing learning objectives were maintained.

Storyboards were created to indicate changes made to content and screens due to the changing from lightpen to voice as the input mode. These consisted primarily of an introduction unit on how to operate the computer using voice, an orientation to the vocabulary, and practice on speaking to the computer.

All screens were reprogrammed and audio recordings were made for feedback purposes and for use with some of the graphics. The module practices and tests were programmed so that the student was required to get 3 out of 5 correct to move to the next instruction.

Evaluation

Pre and posttests, observation data forms, and opinion questionnaires were designed to collect data on the performance of the VBTDS.

Pre and Posttests

A pretest was administered to all 37 subjects prior to taking the instruction. The results of 34 subjects were used for statistical purposes due to incomplete information on the forms. After the instruction, which consisted of 8 modules, a posttest was administered. A comparison between the two tests indicated that 85% of the subjects performed better on the posttest than on the pretest and 15% showed no gain. All of the subjects completed the course.

Observation Data

Observation data was recorded during a field test at Ft. Knox, Kentucky. Thirty-seven soldiers were observed using the system, and data was obtained concerning how well the system recognized the words spoken. It was hoped to obtain a misrecognition factor of no more than 15%. The average recognition factor was 72%. There was a misrecognition factor of 12% and a nonrecognition factor of 16%. The 12% misrecognition factor was better than the 15% standard set at

the beginning of the test; therefore, the voice recognition system performed better than anticipated.

Opinion Survey

An opinion survey was administered to the subjects consisting of thirteen questions that required the subjects to rate aspects of the course and the VBTDS as excellent, very well, good, fair and poor. A Cronbach's Alpha was calculated to estimate the internal consistency of the items and revealed a reliability coefficient of .84. All the ratings obtained were 3.0 or above with an average rating of 4.1 or very good. Descriptive statistics were used to answer the study questions.

Conclusion

When examining the pretest/posttest results, opinion results, and observation results as a whole, the student gained knowledge while completing the lesson, held a high opinion of the course and training delivery system, and exceeded the criteria set for the operation of the voice recognition system. Based on the above information, the following conclusions were drawn in relation to the original research questions.

Question 1. "What is the current state-of-the-art regarding automatic speech recognition and computer instruction?" At the inception of this study (1987), an independent voice recognition board and communications package was not available off-the-shelf. However, Scott Instruments was willing to, and did, develop an independent voice recognition board and speech communications package to be used in a VBTDS. Technology advances have continued to occur, largely through Scott Instruments as a result of this study, and other parallel programs such as the IBM SPHYNX System, which was introduced to the market in 1990. Consequently, the state-of-the-art of this technology has changed since the beginning of this study, partly as a result of this study. Independent voice recognition systems are now a viable choice for computer assisted instruction input.

Question 2. "What are the available options and the minimum requirements for a VBTDS?" An independent voice based training delivery system was developed to present training. The hardware consisted of a PC-AT clone with one megabyte of memory and a forty megabyte hard disk. Two floppy disk drives were included, one five and a quarter and one three and one-half inch. There was a multiscan color monitor with supporting EGA graphics card. The speech recognition system was a single-board, signal processor with

onboard memory for vocabularies up to 50 words. There was a Visage 1800E digital audio board for audio recordings. There was a microphone for speech input. The software supporting the system was MS DOS 3.2 and its supporting utilities, the authoring language TenCORE for creating the CAI, speech recognition software for creating, training and modifying vocabularies as well as two programs that interfaced the speech recognition board with the TenCORE authoring system. There was the digital audio software that included two programs that interfaced the audio board with TenCORE. Authoring commands within the lesson accessed stored audio messages. A graphics software, PC Paintbrush, was also included for the inclusion of graphics images in the courseware. In summary, a VBTDS was developed that integrated an authoring language, graphics and audio output software, and voice recognition input software to allow for development and delivery of voice based training on one PC-AT clone computer system.

Question 3. "Is a system using a VBTDS as the primary mode of learning a feasible substitute for current delivery systems used for training soldiers in CEOI as measured by: a) student learning gain, b) student ratings of the system.

The mean student learning gain was 3.18 between the pre and posttest. The mean students' rating on the course was 4, very good. There was a small but insignificant correlation between students' mean gain of knowledge and the operation of the system of +.13 as well as between students' mean gain of knowledge and that the students felt the system should be used for training. There was no significant relationship found between the mean misrecognition of 12 percent and the students' opinion that the VBTDS should definitely be used for training.

The accuracy level of the recognizer was slightly better than the objective set for this study and did not appear to be detrimental to subjects completing the course or to the scores made on the pre and post test. Additional work to increase the recognition level could be done to elevate the recognition level. The subjects as a whole thought the system worked very well and could be used for training.

Recommendations

Several recommendations were made concerning how the conclusions from the study could be implemented by the Army TRADOC community.

1. The courseware could be enhanced to make even better use of the voice arena for input. Much of the visual information queues could be replaced with audio output messages, requiring the student to write down what was heard, just as must be done when in a tactical situation. The use of more audio messages in the narrative could provide ease of learning for slow readers.

2. More help sequences could be added to the instruction to provide detailed explanations of instructional material.

3. Work should be done to improve the accuracy level of the recognizer. Even though the accuracy level was adequate for this delivery, greater accuracy would enhance the use of the system.

4. The use of a VBTDS could be expanded into training as well as testing in other areas in which performance is also voice based.

Concluding Statement

In the course of this study, several areas emerged which are recommended for further study.

1. This study did not allow for comparing the delivery of training by using voice input with any other modes of input such as keyboard, light pen, or touch screen. A follow-up study examining additional modes of input would provide empirical data about voice in relation to other modes.

2. The Army has a standard training delivery system called the Electronic Information Delivery System (EIDS). The VBDTS was developed using an EIDS compatible hardware configuration and required installing the voice recognition board, audio recording and delivery board, and the training software onto the EIDS to test compatibility. A study in which the voice board is integrated with the EIDS hardware would indicate if EIDS could be modified.

3. There could be many applications for the use of voice input in a training arena. Foreign language

teaching would be a natural use of this type of delivery system. Although some lessons have been developed in German, the system did not have an independent voice recognizer. A foreign language lesson should be developed and tested using the VBTDS.

4. One of the subjects used during the validation had a speech disorder. As he worked with the system, he learned to say the words correctly for better recognition. The system could be tested in teaching the correct pronunciation of speech for people with speech disorders.

The quest to develop training that simulates real world requirements could be met by a delivery system such as the one developed in this study. This system allowed for the added dimension of training individuals to use communication just as it would be used in an actual job situation. It could be used to train customer service tasks, air traffic controller tasks, telephone operator tasks, radio communication tasks, and foreign language scenarios, to name just a few. It was hoped that the findings of this study would aid the Army in efforts to present real world simulated training.

Once again, it must be noted that technology advances in the area of voice recognition systems and the appropriate applications of the state-of-the-art of this technology have changed since the beginning of this study. The VBTDS prototype developed during this effort is, even as this study is reported, now entering the market and in the near future will be regarded as an antiquated attempt. Another illustration of the evolution of this technology is the emergence in 1989 of a trade magazine, Voice Processing, devoted to the area of computerized speech capabilities.

In summary, the major objective of this study was to develop an independent voice based training delivery system and sample courseware. Since no documented studies have shared this objective or attempted to determine the developmental steps for creating voice based courseware, this study was exploratory . Although the results from this study were modest, they should useful and may aid the Army and other potential users of this type of training delivery system in efforts to present training to simulate actual job tasks.

The greater importance of this study, however, was the suggestion offered by the study's findings that future research of this type could significantly advance the use of this form of instructional technology.

References

Adler, William (1987). Introduction Remarks, Official Proceedings of Speech Tech '87, 103.

Author Unknown (1988). "Can Computers Answer America's Training Needs?" Instructional Delivery Systems. 10-11, 14.

Beek, Bruno; Neuberg, Edward; and Hodge, David C. (1977). "An Assessment of the Technology of Automatic Speech Recognition for Military Applications." Automatic Speech & Speaker Recognition, New York: IEEE Press, 101-113.

Bergondy, M. (1987). Advanced Instructor Station Design. Defense Technology Information Center, Matriss No. 350921.

Bristow, Geoff (1986). Electronic Speech Recognition: Techniques, Technology, & Applications, New York: McGraw-Hill Book Company.

Butler, F. Cloit (1976). Instructional Systems Development for Vocational and Technical Training, Educational Technology Publications, Englewood Cliffs, New Jersey.

Chambers, Randall M. and Brown, Mark J. (1987). Land-Based Applications of Speech Technology. Official Proceedings of Military Speech Tech '87, 83-91.

Chambers,Randall M.; Maclay,Neal; Gerrits, Jonus S.; and Brown, Mark J. (1987). Speech Technology for Land-Based Systems. Official Proceedings of Speech Tech '87, 35-39.

Cornick, Lisa (1983). Microcomputer Software for Teaching German: An Evaluative Doctoral Dissertation for Syracuse University. ERIC Document ED234752.

Dallman, B. (1986). Intelligent Speech Recognition Systems. Defense Technical Information Center, Matriss No. 45031.

Dickson, W. Patrick (1986). Experimental Software Project: Final Report. Program Report 86-10. ERIC Document ED276400.

Dickson, W. P.; Neal, V. A.; and Gillingham,M. (1984). A low-cost multimedia microcomputer system for educational research and development. Education Technology, 24(8), 20-22.

Doddington, George R. and Schalk, Thomas B. (1981). Speech Recognition: Turning Theory to Practice. IEEE Spectrum, 18, 26-32.

Ehrens, Ron (1988). Improve (Immediate Production Verification) Voice Recognition for Quality Control. Official Proceedings of Speech Tech '88, 131-136.

Emanuelson, Kim (1987). Speech-Driven Factory Control System (A 14-Minute Video Tape). Official Proceedings of Speech Tech '87. 104-105.

Enlisted Career Management Fields and Military Occupational Specialities (AR 611-201). Headquarters, Department of the Army, Washington, D.C., 1987, 32-34.

Gottesman, Kyra (1988). PC-Based Voice Recognition--A New Voice Messaging Solution. Speech Technology, 4, 89-90.

Horn, Corin E. and Scott, Brian L. (1983). Micro-Based Speech Recognition: Instructional Innovation for Handicapped Learners. ERIC Document LED231146.

Instructional Systems Development (1973). Headquarters, Training and Doctrine Command (TRADOC), Ft. Monroe, Virginia.

Klatt, Dennis H. (1974). On the Design of Speech Understanding Systems, Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium, Ed. Raj Reddy, 277-289.

Klaver, Martin (1988). Voice Input for Production Quality and Process Control. Official Proceedings for Speech Tech '88, 137-140.

Kristiansen, D. (1985). Technology Transfer in Armor Training. Defense Technical Information Center, Matriss No. 250265.

Larson, John T.; Gibson, Marcia R.; and Ballentine,Bruce (1988). Voice-Based Training Delivery System: A Total System Approach, The Official Proceedings of Military Speech Tech '88: Voice Applications for Military and Government Agencies, pp. 85-89.

Lea, W. A. (1980). Trends in Speech Recognition. Englewood Cliffs, New Jersey: Prentice-Hall.

Lee, Kai-Fu (1988). On Large-Vocabulary Speaker-Independent Continuous Speech Recognition. Speech Communication 7, North-Holland, 7, 375-379.

Levinson, S. E. and Shipley, K. L. (1980), A Conversational-Mode Airline Information and Reservation System using Speech Input and Output. The Bell System Technical Journal, 59, 119-137.

Mager, Robert F. (1988). Making Instruction Work on Skillboomers, David S. Lake Publishers, Belmont, California.

Mager, Robert F. and Beach, Jr., Kenneth M. (1967). Developing Vocational Instruction, Pitman Learning, Inc., Belmont, California.

Martin, Marcus B. (1977). Applications of Limited Vocabulary Recognition Systems. Cinnaminson, New Jersey: Threshold Technology, Inc.

Martin, Thomas (1976). Practical Applications of Voice Input to Machines. Proceedings of IEEE, 64, 491-501.

McLagan, Patricia A (1978). Helping Others Learn: Designing Programs for Adults, Addison-Wesley Publishing Company, Reading, Pennsylvania.

Nadis, Steve. (1988). A Machine to Talk To. Technology Review, 12-13.

Nelson, Carl F. (1988). Parts-Data "VOICS"tm System. Official Proceedings for Speech Tech '88, 129-130.

Neuberg, E. P. (1974). Philosophy of Speech Recognition. Ft. Meade, Md.: National Security Agency.

Olsón, Eric J. (1987). Voice Recognition for 100% Parts Audit. Official Proceedings for Speech Tech '87, 95.

O'Shaughnessy, Douglas (1986). Speaker Recognition. IEEE ASSP Magazine, 4-16.

Oshika, Beatrice T.; Eldridge,Charles; and Adams, Duane (1987). Voice I/O and Artificial Intelligence. A tutorial presented at AVIOS '87.

Peckham, Jeremy (1988). Talking to Machines, IEEE Review, 385-389.

Promislow, Mark R.; Larson, Nancy L. J.; Guidry, Paul J.; Eisher, Edward L.; and Joost, Michael G. (1987). Concepts in Multi-Modal Communication. Official Proceedings of Speech Tech '87, 181-184.

Rash Jr., Wayne (1989). A Helping Hand. Byte, 129-130.

Richard, G. L. Ed. and others (1982). Workshop on Instructional Features and Instructor/Operator Station Design for Training Systems. ERIC Document ED226713.

Systems Approach to Training (1987). Headquarters, Training and Doctrine Command (TRADOC), Ft. Monroe, Virginia.

Voice Processing Magazine (1989). Information Publishing Corporation, Houston, Texas.

Wen, Samuel S. (1987). Speech Recognition in the Consumer Market Today. Official Proceedings of Speech Tech '87, 31-32.

Woodward, J. R. and Cupples, E. J. (1983). Selected Military Applications of Automatic Speech Recognition Technology. IEEE Communications Magazine, 35-41.

APPENDIX A

Market Survey Questions

Market Survey Questions

This form contains the information obtained during the market survey visits.

1. Name of company and address.
2. Is the recognizer speaker-independent or dependent?
3. Discrete words or connected words?
4. What is the vocabulary size?
5. Is it context sensitive?
6. What applications have been developed?
7. Are there software tools?
8. What is the cost?

Observation Training

Observation Training

Agenda

Purpose of the Project	15 minutes
Importance of the Data	15 minutes
Role of the Observer	15 minutes
How to Collect the Data	15 minutes
Practice Collecting the Data	60 minutes

Student Profile Sheet

STUDENT PROFILE SHEET

Please answer the following questions as accurately as possible. The information obtained from this questionnaire will be held in confidence and will be used only for this validation effort.

1. Date_____
2. Name_____
3. Present rank_____
4. Time in the Army_____(years)_____(months)
5. Present MOS_____
6. Time in MOS_____(years)_____(months)
7. Have you completed BCT? ____yes _____no
8. Have you completed AIT? ____yes _____no
9. Have you completed BNCOC? ____yes _____no
10. Indicate the highest level of education you have by placing a check by the most appropriate answer below.

    _____high school diploma
    _____General Equivalency Diploma(GED)
    _____no high school diploma
    _____1-4 years of college, no degree
    _____College degree(B.S. or B.A.)
    _____Graduate work, no degree
    _____Graduate degree(M.S. or M.A.)
    _____Post graduate work, no degree
    _____Certificate of Advanced Studies
    _____Doctorate
    _____Other_____(provide information)

11. Have you ever had training in the use of the Communication-Electronics Operation Instructions?_____yes _____no
12. If the answer to 16 is yes, when and what?_____
_____

Thank you for completing this form. Please give it to the proctorwhen you are finished

Observation Forms

SAMPLE -- DATA COLLECTION FORM FOR INSTRUCTIONAL UNITS

UNIT	QUESTION	ANSWER	RECOGNITION	NONRECOGNITION	MISRECOGNITION	
A1	WHAT ACTION TO TAKE	B				
A10	YOUR CALL SIGHN AND SUFFIX	URH91				
A11	UNITS CALL SIGN AND SUFFIX	K9C74				
A11A	RADIO FREQUENCY	71.18				
A12	CORRECT CALL TO ENTER NET	B				
A13	IDENTIFY YOUR STATION	REFER TO AI				
A14	AUTHENTICATE MIR	I AUTHENTICATE B				
A15	PROWORDS, ENCODE MESSAGE	GROUPS 4 KHO HPI AKD UHK				
A16	DECODE MESSAGE	A				
A17	ENCRYPT GRID COORDINATE	TUCKWCFKB				
A18	ABBREVIATE CALL SIGN	H91				

OBSERVATION -- DATA COLLECTION FOR CONTROL WORDS

GO	PROCEED	CONTINUE	BACK	OVER	CORRECTION	HELP
R	R	R	R	R	R	R
N	N	N	N	N	N	N
M	M	M	M	M	M	M
R	R	R	R	R	R	R
N	N	N	N	N	N	N
M	M	M	M	M	M	M
R	R	R	R	R	R	R
N	N	N	N	N	N	N
M	M	M	M	M	M	M

Pretest

CEOI PRETEST QUESTIONS

Once you start the test, you must continue until you have completed the situation. Do not stop, even if you know you have missed more than one question.

You will need a copy of the CEOI to take this test. With the yellow cover up, turn to the first page of the CEOI. Remember, you have only 30 minutes to complete all questions.

Write down the following information:

Time period is:	02
You are:	Track 2, 2/B/2-14 CAV SQDN
Contact:	PLT LDR 2/B/1-14 AR BN

QUESTION 1. What is your call sign and suffix? Find your answer in the CEOI.

QUESTION 2. What is the call sign and suffix of the unit you are to call?

QUESTION 3. What is the radio frequency for the unit you want to call?

QUESTION 4. What is the correct initial call to enter the net?

a. Q0M38 This is FOJ76 Entering your net, over

b. QOM38 This is FOJ76 Request permission to enter the net, over

c. Q0M38 This is F0J76 Identify your station, over.

d. Q0M38 This is F0j76 over

QUESTION 5. What is your reply to "Identify your station"? Include prowords but not call signs in your answer.

QUESTION 6. What is the reply if you were asked to "Authenticate HR"? Include prowords but not call signs in your answer.

QUESTION 7. Decode the following message and say the letter of the matching message below:
GROUPS 3 ZEG YDU SJM

The message is :

a. Seized objective red
b. Seized objective black
c. Reached objective red
d. Seized objective green
e. Reached phase line green
f. Reached phase line red
g. Seized objective violet
h. Reached objective violet

QUESTION 8. What is the abbreviated call sign for: C4E74

Posttest

CEOI POSTTEST QUESTIONS

You will need a copy of the CEOI to take this test. With the yellow cover up, turn to the first page of the CEOI. Remember, you have only 30 minutes to complete all questions.

Write down the following information:

Time period is:	01
You are:	Track 4, 3/C/1-14 AR BN
Contact:	PLT SGT 3/C/2-14 CAV SQDN

QUESTION 1. What is your call sign and suffix? Find your answer in the CEOI.

QUESTION 2. What is the call sign and suffix of the unit you are to call?

QUESTION 3. What is the radio frequency for the unit you want to call?

QUESTION 4. What is the correct initial call to enter the net?

a. K9C74 This is U5H91 Entering your net, over

b. K9C74 This is U5H91 Request permission to enter the net, over

c. K9C74 This is U5H91 Identify your station, over.

d. K9C74 This is U5H91 over

QUESTION 5. What is your reply to "Identify your station"? Include prowords but not call signs in your answer.

QUESTION 6. What is the reply if you were asked to "Authenticate HR"? Include prowords but not call signs in your answer.

QUESTION 7. Decode the following message and say the letter of the matching message below:
GROUPS 3 ZEG YDU KLM
The message is :

a. Seized phase line red
b. Seized phase line blue
c. Reached objective red
d. Seized objective blue
e. Reached phase line red
f. Reached phase line blue
g. Seized objective red
h. Reached objective blue

QUESTION 8. What is the abbreviated call sign for: U5H91

Trainee Opinion Survey for the CEOI Lesson Using a Voice Recognition System

NAME_____

TRAINEE OPINION SURVEY
FOR THE CEOI LESSON USING A VOICE RECOGNITION SYSTEM

You have just completed a computer assisted instruction lesson on the CEOI. Please indicate your feelings about the course by answering the questions below. Your assessment will assist us in fielding a course that will meet the requirements of future trainees.

Please circle the number that best indicates your feelings about the course. The numerical scale is set up so that a 5 is the highest rating you can give the question and a 1 is the lowest rating.

5=excellent, 4=very well, 3=fair, 2=very poor, and 1=not at all

How well did the course meet your expectations?	5	4	3	2	1
How well was the instruction organized?	5	4	3	2	1
How clearly were the objectives stated?	5	4	3	2	1
How well did the instruction cover what the objectives stated?	5	4	3	2	1
How well did the examples and practices help you to understand the lesson material?	5	4	3	2	1
How well did the pictures and illustrations help you to learn the lesson material?	5	4	3	2	1

Name_____

How well did the test match what you were taught in the lesson?	5	4	3	2	1
How well did the voice recognition system operate for you?	5	4	3	2	1
How well did the introduction orient you to operating the computer with your voice?	5	4	3	2	1
How would you rate the voice recognition as a training application for this type of lesson material?	5	4	3	2	1
How would you rate your knowledge of this material prior to taking this lesson?	5	4	3	2	1
How would you rate your knowledge of this material after taking this lesson?	5	4	3	2	1

How would you rate the level of difficulty of this material?	too high 5	4	about right 3	2	too low 1

APPENDIX B

Market Analysis of Voice Input/Output

MARKET ANALYSIS OF VOICE INPUT/OUTPUT

I. Statement of the Problem. Training requirements, particularly in the military environment necessitate customized instruction, with as much hands-on experience and one-on-one interaction as possible. Unfortunately, in many instances, training resources do not permit this. To determine the feasibility desirability of improving the trainer-student ratio with computer aided instruction that incorporates voice input and output, EER Systems performed a market analysis of computer voice input/output technology. The two applications of particular interest to the Army are Russian linguist training and armor crew training (fire, move, and communicate commands). The goal is to identify a training technology, including Voice I/O, that may be incorporated into a training delivery system.

Speech technology is a synthesis of linguistics, computer science, and artificial intelligence (AI). To emulate communication with apparent naturalness, researchers attempt to understand how people speak and hear. The researchers must also be familiar with the computer technology (chips, boards, and algorithms) used to generate and accept voice I/O. Finally, they must be able to apply breakthroughs in AI that deal with natural language programming and rapid search techniques for transaction parsing. Responsiveness is a critical element of an effective training delivery system.

Three components of voice input/output are voice recognition, voice synthesis, and voice digitization. The simplest by far is voice digitization in which voice signals are recorded, stored, retrieved, and played when needed. The stored signal can be changed to modify rate of speech, pitch, inflection, etc. An example of this technology is the TI Talker, made by Texas Instruments. A major drawback to voice digitization is that every spoken utterance must have been previously recorded. The size of such a resulting vocabulary is limited by the availability of data storage.

Voice synthesis is the technique in which parts of words (phonemes) are output in combinations to form words and phrases. In the English Language, there are only about forty phonemes. Although this minimizes the data that must be stored, voice synthesis requires sophisticated software and significantly more processing capability than digitization. A significant benefit of voice synthesis is no speech has to be recorded. A drawback is a perceived monotony of voice because of the lack of inflection by the speaker. The most prevalent application of voice synthesis is in text to speech applications.

Voice recognition is the most-difficult, least-developed aspect of voice I/O. Typically, recognition is performed by capturing speech and comparing the speaker's wave pattern against a preformatted pattern. When a pattern is matched, the computer recognizes the word associated with that pattern. Recently, techniques have been developed to decompose speech into

phonemes rather than to attempt matching. This technique requires less storage but like voice synthesis, demands more sophisticated software. Given the advanced state of the art in voice output (both digitization and synthesis), the remainder of this analysis will be restricted to the investigation of computer voice recognition systems, the critical feature of voice input/output in a training application.

II. Evaluation Criteria.

A. Speaker Dependence/ Independence.

The most critical feature of a voice recognition system in a training environment is speaker (person) dependence/independence. A dependent system is one in which each speaker must prerecord voice patterns to form templates, or "train the system". It is extremely difficult for computers to match the voice patterns of various speakers, given the many differences in tone, accent, inflection, and rate of speech. There are dependent systems that the same speaker system had to be retrained because of a change in their speech due to a cold or just the passage of time.

Speaker-independent systems typically use composites of many speakers as generic templates. This requires additional data storage as well as significantly more front-end processing to accumulate the data and form the generic patterns. For this reason, independent systems have much smaller vocabularies than dependent systems. In spite of the smaller vocabulary, a training system should be speaker independent; this avoids the training time that is required for each trainee to "train the system". Although not desireable, it may be acceptable to have a recognizer that adapts the trainer's voice through a short "introductory" session. This is possible with some extant systems, such as that of Scott Instruments. A new technique that decomposes speech into phonemes instead of matching templates may provide great advances in speaker independence. Unfortunately, these systems are still at the prototype stage and are not available for wide application as yet.

B. Words or Speech

Another critical feature in a voice recognition system is the ability of the system to recognize descrete words or connected speech. Discrete words may have different pronunciations, depending on the placement (context) of the word within a sentence, and on the region from which the individual speaker originates. Further, people tend to slur words together (coarticulation) and sometimes drop sounds, such as the "g" in "ing". Human listeners mentally complete the syllables and judge meanings from syntax and context, but researchers still do not understand fully how this process works, and, therefore, cannot duplicate the capability in computer software. Obviously, discrete speech in which every word has a distinct start and stop, and in which every word is pronounced with only a slight deviation, is easier to recognize with guaranteed accuracy than conversational, continuous speech. Computer engineering technology provides the processing power to recognize connected speech with pauses as small as 300 microseconds, if each word is enunciated distinctly. In a

training delivery system, connected speech recognition is optimum; however a discrete speech recognition system, which can substitute short phrases for words, would probably suffice.

C. Vocabulary Size

Another factor in evaluating voice recognition systems is the vocabulary size. A system's vocabulary is based upon the application, since the application ultimately will determine the effectiveness of the training, based on the time and trouble invested to enter the vocabulary as well as the storage required to maintain it. In addition, searching through a large number of templates requires more time and thus slows responsiveness. In a speech-to-text dictation system, a vocabulary of several thousand words is necessary. However, at least one expert believes a typical application will run effectively with about 200 words. The vocabulary required for armor crew training is certainly small, probably less than 100 words.

D. Context Sensitivity

Development of a grammar that anticipates the available matches based upon the context, would improve response time for large vocabularies In such a context-sensitive system, the recognizer knows that only a limited number of words can follow a given recognized word and, therefore, only searches through those templates instead of the entire database. Use of AI search techniques for grammar construction can make a significant impact on response time, even with small vocabularies, and is recommended for training application.

E. Application

Another consideration in the evaluation should be the possible applications for each system. Speech technology researchers indicate there is no best system for all applications. The application determines which recognizer is most useful given environment, speakers, etc. A voice-driven Computer-aided instruction (CAI) package or defense training system simulator, would be of particular interest for this training application.

F. Development Tools

A voice recognizer system (hardware) does not work in isolation. There must be software to execute the commands and communicate with existing software. If the recognizer does not come with a library of software tools, the applications of voice recognition systems will require much more sophistication (computer knowledge) on the part of the user.

III. Methodology. The examination of voice I/O was conducted through a review of the speech technology literature, attendance at voice input/output conferences, visits for on-site demonstrations, and discussion with experts in the field. The literature consisted of recent conference proceedings, textbooks, magazine articles, journal publications, and product brochures. The conferences attended were the American Voice I/O

Systems (AVIOS) Applications Conference, and Military Speech Tech '87. The examination included discussions with scientists from the Army Research Institute, the Naval Postgraduate School, the Defense Language Institute, and the Air Force Armstrong Medical Research Laboratory in an effort to identify DOD specific research in voice I/O. In all, sixteen systems were evaluated. Seven systems, with direct application to training, were selected for visits on-site, where a closer examination of the system's capabilities could be made.

IV. Findings. A summary of the findings is presented in Table 1. The only speaker-independent system that we were able to identify is produced by Voice Control Systems (VCS). The system has a preestablished vocabulary of discrete words or phrases. VCS develops templates for each word by sampling 500 voices according to a formula it has established to eliminate differences attributable to age, sex, or accent. The standard vocabulary consists of thirty-six words, such as the digits (0-9), yes, no, and directions (up, down, left, right). VCS can develop templates for other words by the same procedure but charges $2500 per word. Speaker independence gives the VCS recognizer great flexibility. The system can be used in security systems, where any voice can activate a computer system that will report a fire, or summon an ambulance. It can also be used to control automatic wheelchairs and to counter the influence of stress on voices in emergency situations.

Scott Instruments (SI) offers the only system that has been used in language training. Scott's Voice-Based Learning System (VBLS) can evaluate student pronunciation and provide corrective feedback. A prototype system is in use at the National Cryptologic School at Ft Meade, MD. Also, Texas public schools are using VBLS to teach German. The recognizer is speaker dependent, but has connected speech capability for a vocabulary of about 200 words. Other applications include inventory control and aid to the handicapped.

The Defense Communications Division of ITT has a speaker dependent, continuous speech recognizer with speech synthesis capability. The Naval Ocean Systems Command uses the ITT system to train air traffic controllers to use proper syntax and vocabulary. Under a separate effort, ITT is working with the Army to put recognizers in attack helicopters. The vocabulary is about 2000 words and is defined by the user, who also defines the grammar.

The most aggressively marketed recognizer is sold by Dragon Systems. The thrust of Dragon's current research is to create a true speech-to-text system with a 20,000 word vocabulary. The present Dragon System is a speaker-dependent, discrete speech recognizer with a 5000-word vocabulary. The system is context sensitive, with a user defined grammar. Harvard University has used it as an aid to teach English as a second language. Other uses have a more commercial nature such as keyboard replacement, dictation, and inventory control.

Voice Connection sells a recognizer that is principally a keyboard replacement. However, the Data Ho Company developed software to use the

system for word processing in any of four languages. By means of efficient table lookup and an IBM graphics package, Mr. Ho can speak to his PC through the Voice Connection system to manipulate Chinese, Japanese, Russian, or Korean characters. The software is flexible enough to allow the speaker to select spoken and printed languages for a limited translation capability. Unfortunately, the system is speaker dependent and will only accept discrete speech. Therefore, it is not very suitable for a training environment.

Texas Instruments is a leader in speech technology, but does not dominate the end market for voice recognizers. Typically, TI systems are marketed by value-added vendors in phone systems, keyboard tools, or toys. The recognizer is speaker dependent, with a vocabulary of fifty words and a connected speech capability. TI products are among the few systems that are not context sensitive, but they do allow for multiple vocabularies that are easily interchangeable.

One of the most widely used systems was developed by VOTAN. The company offers speaker-dependent recognizers capable of discrete or connected speech. The user develops the 300-word vocabulary and the syntax with very friendly vendor software. There is also a library of voice routines that can be called from user-developed programs. These routines make voice I/O as simple as a DBMS does a database query. The VOTAN system is currently used by Delta airlines to sort mail packages by speech. The human sorter controls a conveyor system by announcing city and weight codes for packages as they pass. The voice system is faster and more accurate than keyboard input. VOTAN also provides the recognizer that SCI Systems incorporates in their speech control unit for aircraft and space shuttles.

There is a correlation between the cost of a system and its flexibility/applicability. The state of the art in computer technology allows vendors to sell hardware relatively cheaply. Computer software, on the other hand, involves much more human creativity and thus is much more expensive. Those systems that are principally a board for a personal computer are understandably cheaper than the systems that provide software to develop the vocabulary syntax or to integrate a voice with existing applications. Note that any software not delivered with the system will have to be purchased or developed separately.

V. Conclusions and Recommendations. Manufacturers and developers have focused upon, and continue to focus upon, the hardware and software aspects of the recognition systems themselves. However, the overall effectiveness of a training delivery system is ultimately determined by comparing the investment, in terms of personnel resources and costs, with the benefits that the system produces. With regard to a training delivery system that incorporates voice recognition technology, the resource investment includes front-end development for the training scenario as well as any front-end trainee requirements, such as training the voice system to recognize the trainee's voice patterns.

The most effective training system is one that provides sufficient software tools for the developers to easily construct and refine a training

scenario, a voice-recognition system that is completely speaker independent, and a system that interprets conversational or connected speech. This type of system does not now exist in any integrated form. The major void that remains is the development of the speaker-independent system that recognizes conversational speech. Although each of the necessary components is available in isolation, research continues to explore the integration and interfaces required. The systems available today will meet some training requirements. However, they require a compromise between training effectiveness and front-end investments.

Voice Control System, Incorporated, manufactures a completely speaker independent, voice-recognition system. It will recognize any speaker with an advertised accuracy rate of better than 98%, provided the speaker is restricted to discrete words or short phrases. The cost of this system is not prohibitive if the vocabulary size is small. Off-the-shelf vocabularies, such as those required for dialless telephones, plant security, or wheel-chair control are available for approximately $500. The hardware cost is $1295. These applications have vocabularies of approximately 50 discrete words and the engineering costs associated with forming the voice templates were approximately $2500 per word or $50,000. The projected multiple sales for these applications reduce the cost to the figure stated above, $500. A feasible application, one that requires only discrete words such as for maintenance training, could be developed for under $100,000. The cost effectiveness would be established by the number of systems required and the static nature of the vocabulary. The major advantage to this system is the elimination of any time required for the trainee to "train the system".

Scott Instruments, Incorporated, has developed a speaker-dependent system that will recognize connected speech with a high degree of reliability. Each user must train the systems to his or her voice, which is a drawback. However, Scott has developed a CAI software system that provides a developer with a tool to incorporate a quasi-training scenario to have the user train their voice. This type of system is effective in language training, in which the initial scenario has the speaker pronounce given words so that the system may store the voice patterns for reference in subsequent training. The hardware is external to the computer and costs approximately $10,000. However, Scott is in the process of reconfiguring their system onto a board that will fit into a personal computer, thus reducing the cost to less than $1000.

There are two major advantages to this system. First, the system recognizes connected, near-conversational speech. Second, Scott is the only manufacturer that appears to understand the necessity of having software tools that help developers in creating a training scenario. Though limited in scope and maturity, Scott's CAI system could be enhanced or even replaced by a mature PC-based, off-the-shelf CAI system that has a RS232 interface.

In summary, we recommend a closer examination of the two systems described above. Though neither of the systems meets the total requirements of a generic training delivery system that incorporates voice

input/output, either system could be utilized effectively to meet a customized training application.

APPENDIX C

VITA

VITA

MARCIA R. GIBSON

Personal Information

Date of Birth:	10 February 1950 Long Beach, California
Home Address:	5518 Pebble Lane Norfolk, Virginia 23502
Family:	Paul, my husband is employed at Norfolk Naval Shipyard. We have three sons: Steven, Thomas and Brian

Education

1989 Certificate of Advanced Graduate Studies
Virginia Polytechnic Institute and State University

1982 M.S. in Adult Education
Virginia Polytechnic Institute and State University

1972 B.S. in Education concentration in Spanish
Old Dominion University

Experience

1987 EER Systems
Senior Training Analyst/Program Manager

1985 Norfolk Naval Shipyard
Education Specialist

1984 U.S. Army Training Support Center
Education Specialist/Personnel Psychologist

1983 U.S. Army Aviation Logistics School
Education Specialist

1981 U.S. Army Transportation School
Education Specialist

1972 Old Dominion University, Bureau of Conferences and Institutes, School of Continuing Studies
Assistant Director/Instructor

Membership

National Society for Performance and Instruction
Tidewater Chapter of National Society for Performance and Instruction
Southeastern Virginia Chapter of American Society for Training and Development

Honors

1986 Recipient of a Sustained Superior Performance Award from the U.S. Army Training and Support Center

1987 Receipient of a Sustained Superior Performance Award from the U.S. Navy-Norfolk Naval Shipyard

1987 Recipient of a Graduate Assistantship from Graduate Studies at Virginia Polytechnic Institute and State University

Committees

Co-chair, Membership Committee, Tidewater Chapter-National Society for Performance and Instruction, 1989 to Present

President, Hampton Circle of the Kings Daughters Hospital, 1987-88

Co-founder and executive board member, Hampton Roads Womens Network, 1980

Numerous professional assignments commensurate with positions held at various companies

Publications

"Voice-Based Training Delivery System: A total System Approach." 1988 Military Speech Tech Proceedings. Media Dimensions, New York, New York, 1988.

"Fundamental Practices of Curriculum Development." Shipyard Instructional Design Center-Atlantic, Norfolk Naval Shipyard, Portsmouth, Va., 1986.

"Motivational Strategies for Curriculum Development." Shipyard Instructional Design Center-Atlantic, Norfolk Naval Shipyard, Portsmouth, Va., 1986.

"Managerial Awareness: Power and Conflict." National Association of Academic Women, Deans and Counselors Annual Meeting. Washington, D.C., 1979.

"Managerial Awareness: Videotape Presentation." ASTD Region IV Conference, Williamsburg, Va., 1979.

Marcia R. Gibson

Marcia R. Gibson May, 1990