Human Factors Improvement in Simulator Instructor Interfaces Using Speech Recognition

Randy Saunders
Training and Control Systems Division
Hughes Simulation Systems, Inc.
West Covina, California

2Lt Clayton Perce
2111 Communications Squadron
Kelly AFB, Texas

ABSTRACT

Current simulator systems utilize keyboard and touch screen interfaces for instructor input. This project pursued speech recognition as an interface alternative to enhance instructor mobility and reduce trainer interaction time. A prototype system was built and a number of human factors evaluations were made. Results demonstrate areas in which improvement can be obtained with no net system cost impact.

INTRODUCTION

The computerized heart of a modern training simulator handles interfaces with a number of sources. The student's interaction with the simulator cockpit and/or other representations of weapon system controls receives considerable human factors benefits from the extensive analysis performed during the design of the weapon system. Attention to detail in making the student experience as close to the real situation as possible is the hallmark of any major simulator company. Computer interfaces from other simulator subsystems frequently form the baseline between the simulator manufacturer and subcontractors building simulator assemblies. A good engineering analysis in this area is important to successful contract execution, and must receive high emphasis during the initial months of a program. The evaluation of new interface technologies is essential for user interaction to maintain pace with system complexity.

Our paper describes the results of a research program sponsored by the Training and Control Systems Division (T&CSD) of Hughes Simulation Systems, Inc. and performed by a group of students at Harvey Mudd College (HMC). The program objective was to measure the human factors impact of using a speech interaction system versus the current instructor to simulator interface. The approach was developed by T&CSD and the HMC team was responsible for constructing the prototype and testing different instructor interface ideas using student volunteers.

The instructor interface requirements baseline tested in the program was taken from the simulator for F-16C Communication/Navigation/ ECM systems maintenance designed and built by T&CSD. This device was chosen to provide a reference point reflecting the current instructor interface state-of-the-practice against which to compare the new technology. The initial delivery of this F-16C simulator to the USAF was in 1988.

We begin this paper by defining the basic terms and concepts of speech recognition. The instructor interface of the F-16C simulator before and after the conversion to speech control is then discussed. Finally, the results of experimental trials are presented.

SPEECH RECOGNITION CONCEPTS

Two important parameters define the strategy used by speech recognition systems. A system may be either "speaker-independent" or "speaker-dependent" based on how the phrases to be recognized are defined. Simultaneously, a system may be either "continuous" or "discrete" based on how the speech input is monitored. Each choice of parameters has advantages and disadvantages.

Speaker-independent vs Speaker-dependent

A speaker-independent system uses rules and sound patterns developed during its construction to match sounds with their corresponding words. In a speaker-dependent system, a period of training is used to digitize the voice of each user for each word. The advantage of a speaker-independent system is that it can be used by any number of people without any time and storage spent on training. The disadvantages of such a system are that the users must learn to speak in the way the system has been programmed to recognize and the recurring hardware cost is more than twice as much as an comparable speaker-dependent system. The advantage of a speaker-dependent system, in addition to cost, is that the user is free to pronounce the terms in a natural way. Such a system can support non-English languages with no hardware or software modification, a significant factor in support of foreign training requirements. The disadvantage of a speaker-dependent system is that the user must have a floppy disk or equivalent medium to carry their voice training to the system before it can be used.

Continuous vs Discrete

A continuous speech recognition system monitors the user at all times. The words that the system understands as commands cannot be used in conversation because the system recognises the words whenever they occur. A discrete listening system uses an external switch to tell it when to start listening, such as a push-to-talk switch. Once triggered, such a system listens until the speaker pauses or until a predefined maximum word time passes. The advantage of a continuous system is that the user does not have to use additional switches. The disadvantages of such a system is that the instructor cannot talk to the student using words that the computer understands and that system cost is increased by more than double. This means that the use of a continuous voice system in a simulator environment requires the instructor to use two different sets of terms when talking to the computer and the student. This problem makes a continuous system inappropriate for this application. The advantage of a discrete system is that the memory and processing requirements are much less, allowing a larger vocabulary to be used. The disadvantage is that the terms used must be chosen so that they can be spoken within the time window, typically 2 to 3 seconds.

SIMULATOR ARCHITECTURES

The basic architecture of a simulator such as the one used in this study is shown in Figure 1. The system has five major components:

1) Simulated Cockpit - Contains the controls and displays used by the student as well as the central computer system that runs the aircraft simulation.
2) Support Equipment Panel - Contains the controls and support equipment mockups used by the student for out-of-cockpit activities.
3) Student Interface Station - Permits the student to make control inputs to the trainer, such as signing on, opening doors, and other actions not supported by the cockpit or support panel.
4) Instructor Interface Station - Permits the instructor to select the problem to be presented to the student and to control other attributes of the training situation.
5) Communication Bus - Allows the computers in the other four components to communicate with each other.

This architecture is equally applicable to aircrew training situations in which the support panel is not present, as a visual or motion subsystem could use the same kind of bus interface.

In this kind of a simulator architecture, conversion to voice interaction for the instructor can be done at low risk because the changes are localized in the instructor station. This is important in the design to improve user interaction without impacting system cost.

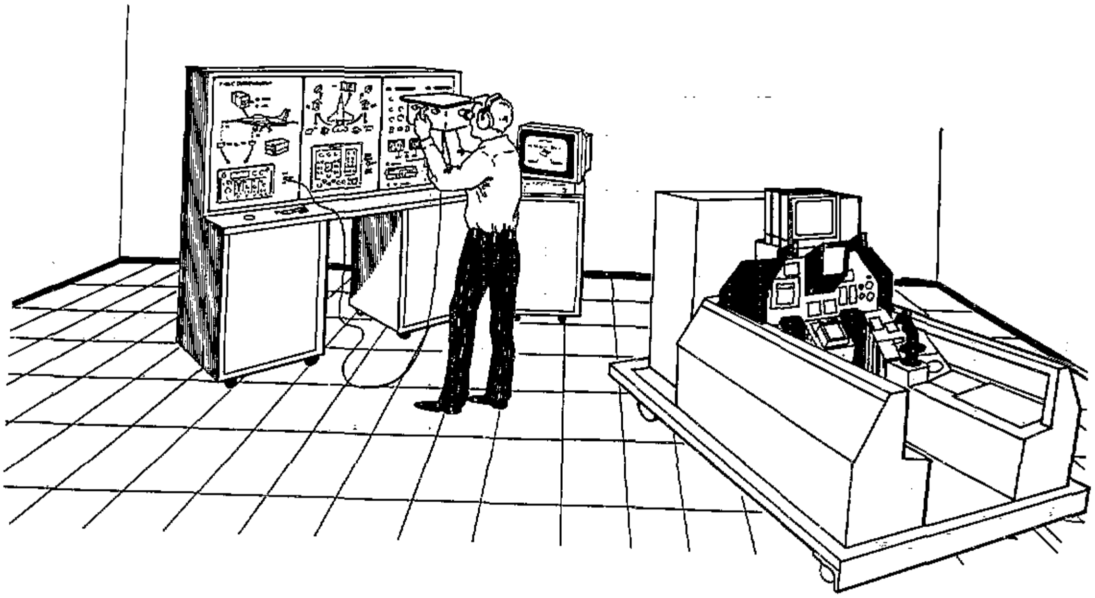


Figure 1. F-16C Communication/Navigation/ECM Simulator

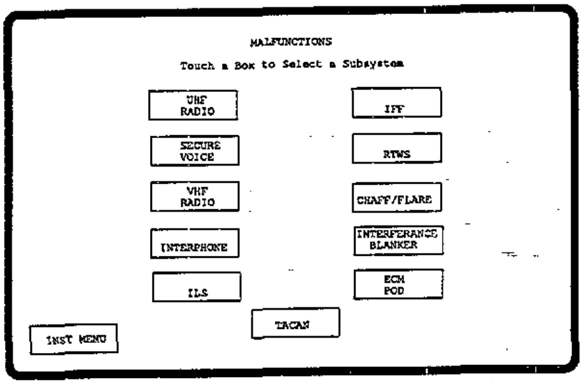


Figure 2. First-Level Malfunction Menu

F-16C Comm/Nav/ECM Simulator Design

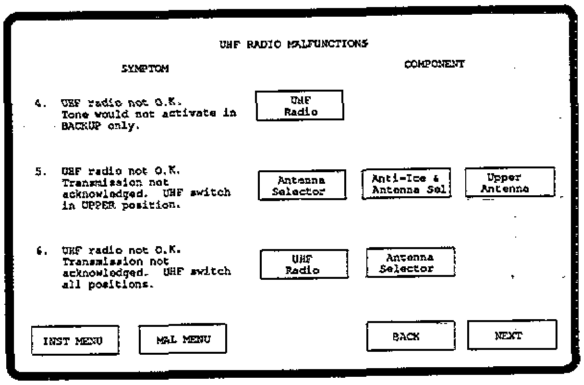
The F-16C Communication/Navigation/ECM simulator has a simple interface system, with a single computer providing both the instructor and student interfaces. The computer uses an IR touch screen in front of a video monitor to detect user inputs. The instructor interface uses a three-level hierarchy to identify malfunctions. At the top level, the menu shown in Figure 2 allows the instructor to select the subsystem containing the malfunction. The second level uses a display like that shown in Figure 3 to allow the instructor to choose the malfunction symptom. Due to display space constraints, multiple pages may be accessed via the NEXT and BACK touch targets may be required to locate the desired system. At the third level, the instructor touches one of the boxes to indicate the malfunctioning component that will be simulated to cause the chosen malfunction.

The advantages of this approach are: (1) the instructor need not remember any codes identifying malfunctions or their causes; (2) the instructor can read the symptom of the failure, without having to consult any notes on the subject; and (3) because the malfunction text is always available, instructor training required in malfunction selection is minimal. Disadvantages of this approach are: (1) a minimum of two touches must be made, with an average of 2.62; and (2) considerable text must be read to determine where to touch the screen.

The system performs adequately, but does not take advantage of the fact that the instructor typically does not require reminders of how the aircraft operates. In fact, the instructor will most likely know the subsystem, symptom, and component necessary to present the desired training situation. Typical means of having this information directly entered, such as via keyboard, require typing skills or extensive code memorization which, have a negative impact on instructor familiarization time and performance.

Voice Interactive Design

The addition of speech recognition to the F-16C simulator would require only the addition of a speech recognition card and microphone to the control console. The instructor would take the microphone and enter the names of the subsystem, symptom, and malfunctioning component directly. Working from stored patterns of the instructor's speech, the system would then identify the items that the instructor would have chosen from the menus in the touch screen system. Advantages of this alternative are: (1) no need to share the touch screen terminal with the student, and (2) quicker entry of selections. The disadvantage is that the instructor must know which subsystem and component is responsible for the malfunction symptoms to be presented or take time to look them up on a reference card.



4. UHF radio not O.K.
Tone would not activate in
BACKUP only.

5. UHF radio not O.K.
Transmission not
acknowledged. UHF switch
in UPPER position.

6. UHF radio not O.K.
Transmission not
acknowledged. UHF switch
all positions.

Figure 3. Second-Level Malfunction Menu

The objective of the program was to provide a proof of concept for this voice interactive approach.

Prototype Development

Prototype development and testing was a four-phase process:

1) Hardware Selection - Available commercial off-the-shelf speech recognition cards were evaluated to determine their applicability to the simulator environment.
2) Language Design - The vocabulary required to specify the simulator's malfunctions was constructed to maximize recognition efficiency.
3) Experimental Testing - The prototype hardware and language were tested on a number of experimental student subjects to determine their effectiveness.
4) System Refinement - Changes to the language to improve performance based on the experimental results were made and tested.

The result was a prototype that supports the F-16C Comm/Nav/ECM simulator instructor interface for selecting malfunctions. The prototype was then used to perform experiments to evaluate its effectiveness over the current design.

Hardware Selection

The HMC team found five manufacturers of cards compatible with the IBM-PC bus used in the instructor station computer. The hardware chosen for the prototype was of the speaker-dependent, discrete listening type. With a 400 word vocabulary, the unit cost of the Dragon Systems VoiceScribe 400 card was around $800.

Another hardware component is the microphone. Both noise-cancelling and wireless microphones were tested and found to achieve adequate results. Wireless microphones have the additional system feature of complete instructor freedom of motion, but at some cost to achieve equivalent noise immunity.

These speech recognition system hardware costs are about twice those of a system that uses an ordinary computer terminal and keyboard. They are equivalent to the cost of a graphics card and monitor suitable for presenting menus for selection with a mouse or function keys. The cost of a touch screen monitor and video overlay graphics board such as that used in the current design is more than double that of this speech system cost. Speech recognition is no more expensive than other user-friendly interaction technologies, and can be traded off during design based solely on technical merits.

Language Design

The most significant driver on system performance was the design of the vocabulary to be recognized by the system. The basic design was patterned after the three-level menu system already in use on the trainer.

Experimental Testing

The parameter most significant to the performance of a speech recognition system is the recognition error rate. This rate measures the percentage of user utterances which are either incorrectly recognized or not recognized at all. These errors require the instructor to repeat the command, significantly increasing the time required to enter the overall selection.

SPEECH INTERFACE TESTING

For each of the experimental trials, two sets of subjects were tested. Their relative error rates of these subjects provide a good measure of the impact of the test topic on recognition error rate. Groups of about 15 test subjects were taken from a pool of 39 male and 15 female college students age 18 to 21. In this test group, no significant age-related or gender-related impacts on error rate were observed. Test topics were:

1) Speaker Volume - The student may overhear the instructor's commands to the simulator if not wearing a headset. To minimize the chance of this, the instructor could speak more softly into the system. Besides lower volume, whispering reduces the spectrum of the voice and therefore increased the average recognition error rate from 5.6% to 8.1%.
2) Microphone Design - The design of the microphone also has significant impact on system performance. Use of a noise-cancelling microphone reduced the error rate from 5.5% to 3.5% in an environment containing significant talking as background noise.
3) Background Noise Sensitivity - The type of background noise has two effects on recognition error rate. First, the level of noise impacts the basic recognition of the system. Second, differences in noise present during training and those present during use is a factor. Tests were performed in three noise environments: quiet, talking, and computer fan noise. Results are shown in Table 1.

Table 1. Noise Sensitivity Results

Error Rate

Testing Environment	Training Environment		
	quiet	talking	computer
quiet	1.9%	6.1%	7.8%
talking	13.6%	3.5%	3.7%
computer	18.5%	4.0%	3.6%

Two interesting results can be seen here: that the training noise environment is a very significant factor in correct recognition and that the presence of noise does not significantly impact the system performance if appropriate training is done. A side result of this testing was a list of areas in which the language design could be improved to reduce the error rate.

Speech Recognition Refinement

Language design was the first area addressed in improving system performance. The following changes were made:

1) Phrase Minimums - Just as phrases that are too long overflow the card's sampling capabilities, phrases that are too short have poor performance. Single syllable words with hard consonants such as "quit" and "back" lack a sufficiently rich sound to make good choices.
2) Phrase Similarity - Two phrases with common words can cause problems depending on the placement of the differences. The terms "UHF Radio" and "VHF Radio" were frequently mistaken because their difference is in the first sound, which typically varies more than others from utterance to utterance. System initial gain also impacts this property. On the other hand "resistor assembly A" and "resistor assembly B" were rarely confused because they have a strong difference in their final sounds.
3) Unusual Terms - The popularity of acronyms in military nomenclature also creates problems. Distinctions between an acronyms that are spelled out when read, such as UHF, and those that are pronounced as words, such as TACAN, caused significant errors. This can be attributed to the experimental subject's lack of familiarity with military terms, since they were college students. However, it points out an area to be avoided if possible in the interest of maintaining the most user-compatible interface.

4) Training Strategy - If the users vary their voice over a range of inflections while training, the system develops a more robust pattern for matching utterances. When a monotone was used in training the system became very sensitive, and only monotone speech could be recognized. Varied voice pitches required less than 3% more training and resulted in the ability to recognize the whole spectrum of voice tones.

The result of use of a new vocabulary constructed and trained using these refinements was an error rate of only 1.6%

HUMAN FACTORS COMPARISON

To provide a comparison of the two alternatives, the following three areas were considered:

1) Time to Select a Malfunction - The most critical parameter because changes in this area are felt at every training session.
2) Percent Error in Selection - The percentage of the time an incorrect selection was made in operational trials.
3) Time to Train Users - The initial training time to achieve productivity at 90% of final productivity.

Each area was analyzed by determining the percentage change caused be using the speech alternative. This means that negative changes represent areas where system utilization is improved by the use of speech interaction.

Time to Select a Malfunction

The time to select a malfunction can be parameterized in the following formula:

$$t_{sel} = t_{access} + t_{enter} + t_{verify}$$

where

t_{sel} is the malfunction selection time
t_{access} is the time to access the interface
t_{enter} is the time to enter the selection
t_{verify} is the time to verify the selection

These three parameters were determined for both alternatives.

Access Time. The value of t_{access} represents the time to get to the interface point. With the typical classroom layout of a F-16C simulator using less than a 20 foot square, the walk resulting in a t_{access} of 7 seconds. Operationally, the touch screen will probably be configured for student functions, requiring the instructor to use a key to switch the system over. Including software delays, this will increase the typical t_{access} to 13.5 seconds for the touch screen system. The voice interactive system can be equipped with a stationary microphone and the resulting walk will give a typical t_{access} of 7 seconds. The addition of a wireless microphone eliminates the walk and results in a t_{access} of only 1 second. In a simulator containing a full mockup of a large weapon system, this factor would take on even larger significance.

Entry Time. The value of t_{enter} represents the time to determine the desired malfunction and make the computer aware of it. In the touch screen system, this is the time to read the malfunction symptoms to find the desired one plus the page selection time to move to it plus the time to scan and choose the malfunction cause. In background tests done at T&CSD, a time of 3.5 seconds per touch was observed in experienced users. This gives an average t_{enter} value of 9.2 seconds. With the speech recognition system, the required three phrases could be entered in 5.0 seconds by experienced users.

Verify Time. Both the touch screen and voice systems must provide user feedback indicating when the computer has understood the user input. The value of t_{verify} also includes the time lost in correcting erroneous inputs. The touch screen system provides audible feedback when the screen has been touched, as well as visual feedback highlighting the selection malfunction. The touch screen system has an error rate around 6%, with a retouch time of 1 second, giving a t_{verify} of 0.06 second.

During the evaluation of the speech recognition prototype, three feedback mechanisms were available to users: a) a green light to indicate when the system had understood a phrase; b) a red light to indicate that the system had not understood a phrase; and c) an audible tone to indicate a phrase not understood. User preferences were split nearly evenly among (b) and (c), with about 65% of users desiring one or the other or both. Option (a) was desired by a very small number of users, with the majority feeling that waiting for a light to come on slowed them down. There was no data available to prove this opinion, but user perspectives must be considered in the design of any such system. The typical error rate, using the optimized language, of 1.6% times the reentry time of 1.7 seconds gives a t_{verify} of 0.03 seconds.

The resulting values for t_{sel} are then:

touch screen = 22.76 seconds
speech = 6.03 seconds

Time to Train Instructors

The time to teach instructors how to use the interface must not be confused with the time to teach the definition and meaning of the malfunctions. The time to learn what malfunctions are supported by the simulator and what lessons to teach with them is equal for both alternatives and is thus not included in this analysis.

As a result, for the touch screen alternative, the interface specific training time is only that time required to familiarize the instructor with the controls on the screen. Although no scientific measurements of this time have been made, a value of 10 minutes is probably representative. The speech recognition system requires less familiarization, perhaps only 5 minutes, but the system must be trained to the user's voice. Results of experimental trials at HMC indicate that training to provide the accuracies discussed can be accomplished at a rate of 90 phrases per hour. The F-16C Comm/Nav/ECM vocabulary has 140 unique phrases, and a training regimen of 5 iterations per phrase results in a vocabulary training time of 95 minutes.

The resulting values for total training time are then:

touch screen = 10 minutes
speech = 100 minutes

Human Factors Comparison Summary

The overall effect of these figures must consider the frequency with which new instructors are assigned to the trainer in comparison to the frequency with which they change malfunctions.

No precise field data is available on the frequency with which the parameters of an F-16C trainer are changed, but one change every 15 minutes will be used as a representative value. This means that the 16.6 seconds saved by the speech recognition system represents a change of -1.8%.

The instructor turnover rate is also not known exactly, but the duration of a training course in the area of 80 days can be used as a valid lower bound. This means that the 100 minute cost of the initial instructor training represents a change of +0.3%.

This gives instructor productivity a net increase of 1.5% for a hardware change costing under $1000. Although this is hardly a result indicating that the existing interface should be scrapped, it represents an incremental improvement that can be applied to future simulators.

EXTRAPOLATION TO OTHER SIMULATORS

The purpose of the program was not to find a solution applicable solely to the F-16C system, but to an entire simulator product line. A number of factors indicate that the switch to a speech recognition warrants consideration:

1) Trainer Configuration - One of the most significant changes in the instructor interface with a speech system is that it frees the instructor's hands and eyes to monitor the student while changes are being made. In simulators utilizing full-size mockups, team training activities or where safety concerns exist, the ability to keep the instructor's attention focused on the students may be even more important than the small improvement in productivity.
2) Simulator Complexity - The F-16C simulators are designed to teach specialized maintenance personnel. As a result some of the F-16C weapon system complexity has been reduced through the partitioning of AFSCs. In a simulator providing whole system training, the complexity of the system controls is significantly increased. This causes the menu-based user interface to become more cumbersome as the number of menus needed to present all the alternatives increases. The speech recognition system does not increase in complexity as rapidly, because phrases are repeated more and because the user can have direct access to many more options for each phrase than any menu system can present for each screen.
3) Instructor Activity - The F-16C simulator requires relatively few instructor inputs at infrequent times. A simulator with more parameters under the instructor's control or needing more frequent instructor interaction would be better suited for speech interaction.

Similarly, there are factors that cause a speech recognition system to introduce additional problems:

1) Classified Systems - Use of wireless microphones is obviously not a good idea in a simulator where classified topics are discussed. The use of wired microphones can solve this, but with some additional inconvenience.
2) Single-Function Part-Task Trainers - The devices which have a user interface which is simple enough to dedicate function keys or touch screen areas to each option will continue to be areas where speech recognition does not provide any productivity improvements.
3) Close Instructor/Student Operations - When the instructor and student share a small work area and have ongoing direct dialog (without headsets), there is no easy way for the instructor to break away and change simulator parameters without tipping off the student. A visual-based interface is more easily oriented away from the student's field of vision.

SUMMARY

Speech recognition can be used to reduce instructor interaction times by two thirds. These systems are a very attractive alternative in simulator situations where a significant portion of the instructor's time is devoted to controlling the computer. The impact on simulator design is low and does not represent a significant project risk. There is no cost impact in comparison to other modern means of computer interface, and so speech recognition should be considered in the instructor interface of future simulators.

The team would like to express their thanks to Dr. Gary Poock of the Naval Postgraduate School in Monterey (55K) for significantly improving the students' understanding of the application of speech recognition.

ABOUT THE AUTHORS

RANDY SAUNDERS is a project Technical Director with the Training and Control Systems Division of Hughes Simulation Systems, Inc. He has served as the lead engineer on the development of advanced software development tools for the past 7 years. He has a Master of Science degree in Computer Science from the University of Southern California (1983) and a Master of Engineering degree from Harvey Mudd College (1980).

CLAYTON PERCE is a second lieutenant in the United States Air Force. He is serving as a communications/computer systems officer with the 2111 Communications Squadron at Kelly Air Force Base, Texas. He was formerly the project leader for the Harvey Mudd College research team. This program was conducted as part of the school's Engineering Clinic program, which provides students with an opportunity to work with industry on current research topics. He has a Bachelor of Science degree in Engineering from Harvey Mudd College (1989).