VOICE-AIDED TRAINING

W. S. Meisel
President
Speech Systems Incorporated
Encino, California

ABSTRACT

Computer voice response and computer speech recognition can provide a valuable supplementary training aid for military training systems. For training devices based on media or software which is hard to change, voice can provide supplementary information and allow spoken amendments to course information. For simulators which attempt physical fidelity to a particular defense system, speech provides a medium which can communicate information to and from the student without interfering with the defense system displays or controls.

It is feasible to add computer voice response of telephone quality to most training systems. Such voice response can be entered and changed by simply speaking into a microphone. At least thirty minutes of such speech can be stored and retrieved digitally in a simple hardware implementation.

Speech recognition capability can add a further dimension to a voice training aid, allowing the trainee to make requests or to answer multiple-choice questions.

INTRODUCTION

In this paper, we will look at two major categories of training situations which arise: 1) with simulators or part-task trainers, and 2) with training devices, such as videodisc or computer-aided instruction systems. This paper will discuss where we believe that computer-voice response, i.e., the ability of the computer to communicate with the student by voice, can be an important supplement to these training technologies, making these technologies more effective. It will also discuss how speech-recognition capabilities, i.e., the ability of the student to respond to the computer by voice, can be a valuable supplement to computer voice response in some situations. We will discuss how current state-of-the-art hardware and software can be used to meet the requirements generated by these applications.

There have been more specialized discussions elsewhere of the role of speech technology in training, particularly in situations where the response to be trained is a vocal response and in which the specific syntax and words used are part of the training (1,3-5).

SIMULATORS AND PART-TASK TRAINERS

There is a wide range of training systems being developed for military and other applications. These devices include very complex weapon system trainers. Weapon system trainers simulate in real time the response of the operational weapon system to the trainee's actions. At the other end of the spectrum are more simple familiarization trainers where the device responds to a limited number of console actions to give the trainee experience with a particular set of switches and controls.

Simulators and Computer-Aided Instruction

The major characteristic of such trainers is that they attempt to simulate to varying degrees the actual system for which the training is being undertaken. Thus, the displays, switches, and other means of communication between the trainee and the system are those present in the actual weapon system. If we wish to automate such a system to provide computer-aided instruction as a supplement to the instructor, we are faced with the difficulty that we do not wish to interfere with the fidelity of the student-system interaction. We do not wish for there to appear on the CRT screen of the weapon system a question or a prompt that would not be available in the actual weapon system. Misuse of a simulator/trainer as a computer-aided instruction device could too easily result in confusing the trainee between what he can expect in an operational situation and the training situation. We would therefore like to have such machine-based instruction occur without interference with the operation of the training system.

Having the computer-based instruction be delivered by spoken messages is probably the ideal alternative for this type of system, particularly since the student is used to receiving vocal instructions from the instructor while he is operating the training system.

Let's consider a more specific example. Suppose we wished to teach emergency procedures with a part-task trainer. The system might be such that, when a particular weapon system panel failed, the procedure was to use keyboard entry. Because the major part of the training will be for normal operation, training in the emergency procedures can be a poor use of time for an instructor, particularly if these procedures are complex enough to require many repetitions for familiarization. Yet, no one would argue against the importance of training in these procedures.

An example of the interaction with a speaking computer-aided- instruction (CAI) system might be as follows. The CAI system might say, "Suppose the firing panel has failed. What would you do in order to initialize your missile battery for firing?" The user responds as he would in an operational situation, typing in his response. If the system is implemented such that the CAI system is aware of what response was typed in, the machine can respond to a correct answer by saying, "That is correct," and proceeding to the next step in the instruction. If the answer is incorrect, the CAI system might indicate the nature of the error; for example, it might say "You transposed a C and an A in the command. Please try again," or "You used the command appropriate for a failure of the navigation panel instead of the firing panel. Please try again."

If the CAI system is not set up in such a way that it can read the specific response of the trainee, it may simply state the correct response and ask him if he requires further explanation. This approach requires that the student be able to interact directly with the CAI system to indicate that he or she needs more help or a repetition of the instructions. For the reasons previously discussed, it would be useful if the student's responses could be vocal. It would be useful if he or she could state what the response was and have the computer comment upon it, as if the CAI system were aware of the keys pressed.

Even in a situation where the CAI system is aware of the key presses, there should be a mechanism for the student to communicate with the system to request help, a hint, or more detailed explanation.

Speech Technology Requirements

Let us discuss the voice response and recognition technology which would be desirable for voice-aided training as we have described it.

Voice response. A highly desirable requirement is that the speech be easily intelligible and natural speech. The student is learning a difficult task and it is inappropriate for him to also be learning to understand unnatural speech. We would like the speech to be as if it were coming from a recording.

It would also be useful if the speech response system could produce sounds as well; for example, it might be useful to explain to the student that when he had pressed a particular sequence of buttons, he would hear a particular sound and have the system replicate that sound. Similarly, having different voices would be useful; for example, one could use a female voice for instructions about the CAI procedure and a male voice for the instructional material.

Another requirement is that the vocabulary of the speech response system be essentially unlimited to allow instruction to be designed and changed without artificial constraints.

A third requirement is that there be no significant delay in the beginning of a vocal response, even if the content of that response depends upon the student's actions. If there is a differing response for a correct answer than an incorrect answer, the machine must immediately present the appropriate response. This is also the case if the student requests a review, more detail, or a hint.

We must also require that there be a capability for storing sufficient speech to carry the whole training session. This would probably require that a minimum of fifteen minutes of speech be stored in the speech response system. It is unlikely that more than an hour of speech would be required for any single training session, considering time allowed for student response and for repetition of material. A subsidiary requirement is that the speech/material be changeable by reloading the system in some way from some off-line storage medium, such as tape, so that the course content could be easily changed to another hour of speech.

A highly desirable capability is that the system allow the speech content of the course material to be easily changed. Because speech is such a natural medium in which to teach, the full advantage of the speech response technology will not be realized unless the instructor or course designer can readily change portions of the spoken material. This can be required for a number of valid reasons:

1) The instructor may discover that the students have difficulty with a certain portion of the course and require additional instructions that are not in the present course material.

2) The weapon system itself may change, requiring that the course material be changed to reflect this. (It is particularly advantageous when the course material can compensate for an incorrect response of the weapon

system trainer, due to a late change in the operational weapon system.)

3) An instructor may discover that a certain portion of the instructional material is wrong or confusing.

It would be ideal if the instructor could change the course material simply by indicating the portion of the course to be changed and then simply dictate vocally the change or addition.

A final requirement is that the addition of the voice response have a minimum impact, if any, on the hardware and software of the training system. Given the difficulty of developing training systems, any significant increase in the complexity added to that task by a CAI adjunct would be a severe impediment to the addition of such capability. For this reason, it is probably inadvisable to integrate the voice-response capability into the weapon-system hardware and software.

It is modern design philosophy for computer-based systems to modularize as much as possible to simplify the software development. The implications for the computer-voice response in this context is that it would be appropriate for the device to be a stand-alone device where changes in the spoken material were made independently. The communication with the training system could be through simple identifiers without requiring the training system to store the speech material.

The speech response requirements are summarized in Table I.

Speech Recognition. We have discussed the requirements for the voice-response capability of the CAI system. The requirements on the speech-recognition side are less substantial. Speech-recognition capability in combination with voice response would be very powerful if the speech-recognition capability could distinguish the following commands:

"Repeat" --
The student could use this to request that course material or a question be repeated.

"Hint" or "Help" --
The student could use this to request further information or a hint to the correct response.

"Yes" and "No" --
To allow the student to respond to questions.

"Stop" and "Ready" --
To allow the student to stop the CAI process or to inform that he or she is ready to proceed again after an interruption.

In addition, in specific applications, it might be useful for the student to respond by a series of digits or by other specific responses to inquiries by the voice-response unit.

If these requirements could be met at a price commensurate with the cost of the trainer and justified by the service provided, it is quite likely that the device would provide an effective aid in assisting the instructor in training for certain types of skills.

TRAINING DEVICES

Another type of training which is receiving growing interest in the military and elsewhere is the use of dedicated general-purpose training devices. Such devices include the following: (1) general CAI systems based upon minicomputers or microcomputers, and (2) interactive video systems, both videotape and videodisc. In particular, microcomputer-controlled videodisc systems hold great promise for highly powerful interactive training systems at a moderate cost. A major disadvantage, however, of CAI systems in general, and of videodisc systems in particular, is the difficulty of changing such systems once the program is designed. To change a videodisc, for example, may require creating a new master.

This problem could be minimized through the use of an adjunct voice-response system controlled by the same microcomputer that controls the CAI or videodisc system. Thus, while a given frame is on the screen, the speech-response system could be instructed by the microcomputer to speak a given section of material. The speech material could differ depending on the trainee's response.

The same considerations, with respect to being able to change the material easily, apply to this type of system as they did to the simulators and part-task trainers in the previous section. It is almost an axiom that there will be errors in a given set of courseware no matter how often it is checked before being committed to software or to videodisc; a related postulate is that the course material will become outdated in part as soon as it is finalized. Thus, the voice-response system requirements of Table I are relevant to the present section.

In the area of the supportive speech recognition devices, however, there are different considerations. We do not have the requirement in most cases of physical fidelity. Therefore, there is no difficulty in allowing the trainee to communicate with the CAI system through the use of a keyboard or other means such as a touch-sensitive screen, minimizing the need for speech recognition.

Table I:
Requirements for
Computer-to-Trainee
Voice Response

A. Easy-to-understand speech

B. Easily changed speech material

C. Unconstrained vocabulary and syntax

D. Different spoken material depending on student action, with no significant time delay before response starts

E. At least fifteen minutes of speech, one hour desirable; course material changed by reloading from off-line storage (e.g., tape)

F. Sound effects and different voices possible (useful, but not important)

G. Minimal impact on training system hardware and software

There would be a significant motivation for using a speech-recognition system as a response mechanism if the response of the user required a minimum of familiarization by the trainee with the speech system; that is, if the speech system took advantage of the naturalness of speech to the user as a means of communication, it might make the interaction with the CAI system less intimidating. It would be nice, for example, if the videodisc or CAI system could display a multiple-choice question on the screen and ask the user to repeat the correct response. The response choices would ideally be lengthy phrases with no vocabulary constraints. If the system could distinguish the correct from the incorrect responses, this would be a very natural means of interacting with the students. It would have the added value of having the student repeat the correct answer orally, rather than simply press a button marked A,B,C, or D. It is reasonable to assume that a student would better retain an answer which required oral repetition. This would seem on the face of it to be an easy task for a speech-recognition system, since the phrasing of the multiple-choice answers, particularly the wrong answers, is very much under the control of the designer. It would therefore seem to be easy to select choices which are distinctly different.

In the next section, we discuss the implications for speech-recognition technology of the requirements of this section and the previous section.

IMPLICATIONS FOR SPEECH-RESPONSE AND SPEECH-RECOGNITION TECHNOLOGY

There are two distinct technologies involved in this discussion: 1) voice response, and 2)speech recognition.

Voice Response

There are four major approaches to speech synthesis: (1) analog recording, (2) off-line encoding, (3) phoneme synthesis, and (4) waveform coding. In this section, we will discuss these technologies and the degree to which they can meet the requirements of Table I.

Analog Recording. Because they are not well-suited for allowing interactive responses, analog (tape- or drum-based approaches) are not attractive for voice-interactive systems. However, if one wishes to sacrifice voice interaction, one can use a standard cassette system with tones indicating the next CAI or videodisc segment. For some applications, this may be cost-effective; as a general-purpose interactive speech-response system, it is not.

Off-Line Parameter Encoding. Techniques such as linear predictive coding (LPC) are used to analyze a spoken word or phrase and reduce the storage requirements for storing that word digitally. The encoding is done off-line, one word or phrase at a time, on a system different from the system which synthesizes the speech. The synthesis system can be inexpensive. The resulting speech quality is related to the amount of data reduction produced by the coding, but, in general, the result is easy to understand. This technology is suitable for short responses and material which does not require changes; but the difficulty of changing the speech and the large amount of speech required eliminate this approach from practical consideration for the applications of this paper.

Phoneme synthesis. One can describe a phoneme synthesis system roughly as a device which pronounces each letter of a word which is spelled out. The speech can be entered as a string of "phonemes" (analogous to letters) and will be pronounced as entered. A great deal of effort is required to enter phonemes, pauses, and stresses so that the resulting speech sounds reasonably natural; even with effort, the speech is "robot-sounding," and the listener must expend some effort to understand it. It is feasible to use phoneme synthesis for the training applications of this paper; but this technology places a burden on the trainee in understanding the speech, and changing the material can be difficult and time-consuming. Text-to-speech systems under development may ease this latter difficulty (2). This approach conserves computer memory more than any other approach, so a great amount of material could be stored. Rapid retrieval of any part of the material is easily possible. This could be a relatively low-cost approach to voice-aided training with the disadvantages noted.

Waveform coding. This technology basically stores a replica of the speech waveform in real time as the speech is spoken. Coding techniques used in communications, such as Pulse Code Modulation (PCM) coding, can be used to reduce memory storage requirements. It is theoretically possible to use techniques such as LPC computed in real time to reduce storage even further. With PCM coding, speech data can be stored at about 4,000 bytes per second to create highly intelligible telephone-quality speech. The speech can be generated and changed by simply speaking into a microphone.

Waveform coding satisfies all the requirements we outlined for voice-aided training in Table I. It has the further advantage that it is a well-developed technology. Unfortunately, the large memory requirements make this one of the more costly solutions; one half-hour of speech requires over seven megabytes of storage. On the other hand, Winchester disc drives of ten to twenty megabyte capacity are readily available and becoming increasingly inexpensive. Waveform coding is a full solution to the voice-aided training requirements for voice response in Table I -- although not an inexpensive solution.

Speech Recognition

The simple control words required by the simulator application can be implemented with any commercial isolated word recognition device. Distinguishing words such as "help," "repeat," etc., is not difficult.

An isolated word recognizer is not so suitable for distinguishing multiple-choice questions of essentially unlimited vocabulary. The isolated word recognizers require each word or phrase to be less than two seconds or so; each such word/phrase to be used must be individually spoken several times by the trainee to "train" the recognizer; and a long multiple-choice reponse which is a series of shorter word/phrases requires a distinct pause between word/phrases. Isolated word recognition is a complicated way to solve what, in this case, is a simple problem.

Because of the flexibility we have in choosing the multiple-choice responses, we can choose them to be different in the number of "syllables" (more accurately, in the number of energy pulses). By simply monitoring the energy envelope of the trainee's response, the speech recognition system can distinguish trainee responses which differ in energy pulses. Since no spectral information is required, the system is speaker-independent and requires no training. Since this approach cannot distinguish "yes" from "no," it does not replace isolated word recognizers where distinctions between words of equal syllable count are required.

The problem of counting energy pulses consistently over many users is more difficult than implied here, but has been demonstrated. In particular, adding this type of recognition capability to a microcomputer-based speech response system requires a minimum of additional hardware.

SUMMARY AND CONCLUSIONS

If one wishes to use CAI with training systems which replicate part of a weapons system,, voice response and speech recognition allow the trainee to interact with the CAI system without interfering with the fidelity of the trainer.

For training devices, voice response can allow changes to the training material to be made easily by dictating into a microphone. Speech recognition allows trainees to respond to the system in a manner more comfortable to them than a keyboard and in a way which may improve retention of course material.

It is feasible to have all the key characteristics implied by these applications in a voice-response system with current waveform-coding technology. Because of the memory storage required by this approach, a system with a hard disk drive is required; although such devices are becoming smaller and declining in cost, this approach is relatively costly. The cost, however, is low compared to the cost of simulators, part-task trainers, and their instructors' salaries. The cost is not so low compared with the cost of a single low-cost training device, but perhaps acceptable if a single speech response system serves some ten to twenty training devices.

A possible lower-cost alternative is a text-to-speech system using phoneme synthesis. Currently the quality of the resulting speech is questionable; but research continues.

Speech recognition has been discussed in this paper as a supplement to voice response rather than as an end in itself. This approach is motivated by the realization that it is fairly easy and inexpensive to add both isolated word recognition (using board-level systems) and speaker-independent phrase recognition (syllable-counting) to a microcomputer-based, voice-response system. Except for multi-user environments, the speech response/recognition system should never be talking while it is listening, or vice versa; thus, a single microcomputer can control both voice response and recognition. The microcomputer can also handle communications with the training system through a standard serial interface to minimize any impact on the trainer hardware and software.

In implementing a total system (voice response, two types of speech recognition,

and communications with the trainer computer), one accomplishes more than a cost reduction. The speech recognition and response capabilities complement one another and yield a highly versatile system. For example, the system may interact with the trainee by giving vocal instruction and requesting a multiple-choice response which is interpreted by the syllable-counter. The voice-response system can then ask a question which requires a numerical response; that response can be interpreted with the isolated word recognizer. The isolated word recognizer can accept control words such as "help," "repeat," or "wait."

The technology for effective voice-aided training is available; it has the potential to make computer-aided instruction more practical in many applications.

REFERENCES

1. Breaux, R., M. McCauley, P. Van Hemel, "Engineering Design Guides for Voice Technology in Navy Training Systems," this proceedings.

2. Bassak, Gil (ed.), "Giving Voice to Text," Electronics, February 10, 1981, pp.117-125.

3. Harris, Steve (ed.), "Voice-Interactive Systems: Applications and Payoffs," Proceedings of a Symposium, Dallas, Texas, 13-15 May 1980.

4. McCauley, M., and C.A. Semple, "Precision Approach Radar Training System (PARTS): Training Effectiveness Evaluatiion," NTEC 79-C-0042-1, Naval Training Equipment Center, Orlando, Florida, 1980.

5. Van Hemel, P.E., et al., "Training Implications of Airborne Applications of Automated Speech Recognition Technology," NAVTRAEQUIPCEN 80-D-0155-1, Naval Training Equipment Center, Orlando, Florida, 1980.

BIOGRAPHICAL SKETCH

Dr. William S. Meisel; President, Speech Systems Incorporated; formerly Manager, Computer Sciences Division, Technology Service Corporation; Ph.D., Electrical Engineering, University of Southern California; author of a textbook and over fifty technical publications.