SPEECH SYNTHESIS FOR COMPUTER ASSISTED INSTRUCTION:

THE MISS SYSTEM AND ITS APPLICATIONS

by

William R. Sanders, Gerard V. Benbassat and Robert L. Smith

Institute for Mathematical Studies in the Social Sciences

Stanford University

1 Introduction

The Institute for Mathematical Studies in the Social Sciences at Stanford (IMSSS) has developed a synthesis system, MISS (Microprogrammed Intoned Speech Synthesizer), designed to test the effectiveness of computer-generated speech in the context of complex CAI programs. No one method of computer controlled speech production is completely satisfactory for all the uses of computer-assisted instruction (CAI). The choice of synthesis method is strongly related to the kinds of curriculums and instructional designs that will use speech. We chose to use acoustic modelling by linear predictive coding as the method of synthesis for MISS. (1)

In Section 2 we describe criteria appropriate for organizing the comparison of voice response systems for use with instructional computers. Then we describe the particular requirements imposed by curriculums at IMSSS, review general voice synthesis techniques, and finally discuss our actual choice. In Sections 3 and 4 we outline the hardware and software that have been created to support MISS in operational CAI at Stanford.

In Section 5 we discuss the applications of audio to CAI. We mention previous uses of various synthesis methods, with a view to determining the actual instructional use of the audio in these applications. The permanent viability of audio in CAI will depend, we believe, on showing that an audio component adds significantly to the kind of instruction that is possible. Thus, it is important that audio not merely take the place of a hard-copy manual or of printed text on a computer terminal.

We are attempting to use audio to contribute to the solution of the most elusive problem of classical computer-assisted instruction: how to have a dialog with the student about what he is undertaking to do. Solving this problem involves much more as well: namely, having a good semantic representation of the student's work in the curriculum at hand, and having a facility for processing natural language on the computer.

Previous work with natural language applications to CAI convinces us that the output problem -- how to say something reasonable and informative -- is more crucial to good dialog than being able to understand complex "natural language" input typed by the student. This is particularly true in the context of programs like the set theory instructional program (see [20]). Moreover, it is important to separate for the student what he is doing (constructing a proof, writing a program) from the program's commentary about it. With the exception of foreign language instruction, all of our current applications will use audio as a "meta-language" of informal dialog.

2 Choice of a Speech Synthesis Technique

There are many methods for providing computer controlled speech to a student using a CAI system. In this section we describe criteria appropriate for comparing voice response systems used with instructional computers, then discuss the particular requirements imposed by curriculums at IMSSS, then briefly review general techniques used by voice response systems, and explain IMSSS's choice in terms of our criteria of comparison.

2.1 Criteria for Comparing CAI Speech Systems

We use ten criteria in organizing our comparison of voice response systems for application to CAI. The first three measure what the speech sounds like, the next four measure the system's ability to compose and manipulate speech, and the last three measure system implementation parameters that have a major effect on cost.

(1) We thank Professor Patrick Suppes for his direction of this project. We also thank the many people who talked with us about their research and our project, including B. Atal, J. Olive and L. Rabiner of Bell Laboratories, R. Morris of Carleton University, R. Reddy of Carnegie-Mellon University, J. Allen of MIT, and C. Bush, A. Peterson and B. Widrow of Stanford University. Many IMSSS personnel contributed to this work in various ways, particularly T. Dienstbier, R. Schulz, K. Powell, G. Black, D. Webb, E. Andrew, R. Roberts, and L. Blaine. And most of all we express our sincere gratitude to D. Duck for arousing the authors' interest in manipulated speech. The research reported here was supported by the National Science Foundation under NSF Grant EPP 74-15016, A01.

The intelligibility of a speech system is the ability of that system to elicit the correct identification of utterances from listeners. The intelligibility of a system is measured empirically by playing phonetically balanced lists of words to native listeners. (2). Since it is well known that listeners can often correctly identify highly distorted words, high scores on such tests do not necessarily mean that the system is capable of producing all the typical sounds of the language involved.

Coverage describes whether a system is capable of producing all the sounds of speech. A system may cover all the sounds of speech in all languages, or may cover only those sounds typical of a particular language, or may fail to produce all typical sounds even in a particular language. Coverage can be measured like intelligibility except that the word list must contain pairs of words that differ by a single distinctive sound feature.

Quality is a measure of the fidelity of the speech, that is the precision with which natural, humanlike speech is produced. This is primarily a subjective measure related to whether or not listeners perceive the speech as typical of that produced by a native speaker. However, it is also strongly related to the absence or presence of unnatural noises, such as hisses or clicks, produced with the speech. Such noises might not distract from a short intelligibility test but would affect the willingness of a student to listen attentively to the speech.

The flexibility of accessing utterances is a description of the limitations placed on the production of dynamically composed messages due to the inability to retrieve the speech representation quickly enough to present it to a student. Linear accessing is constrained to the retrieval of only the next recorded message. Random accessing allows the rapid retrieval of any recorded utterance. Locally random accessing allows the retrieval of utterances physically nearby on the storage device, but the time delay to retrieve a distant utterance would be excessive.

The ability to compose sentences compares those systems with only prerecorded messages to those which may compose sentences out of smaller speech segments, such as words.

The ability to compose words compares those systems with only prerecorded words to those which may compose words out of smaller segments, such as morphemes or phonemes.

The ability to manipulate intonation compares how well, if at all, systems can change the intonation of a sentence. (3) A system with such ability could, for example, change a message that would normally be perceived as a declarative statement into one that would be perceived as a question.

The vocabulary storage size can be a major and direct factor in the cost of a synthesis system. Within a particular synthesis technique, increased intelligibility and quality usually cause increased storage size, while the ability to compose sentences, words, or intonation cause decreased storage size.

The data transmission rate determines the cost of speaking to a student physically remote from the instructional computer. To provide a means of comparing analog and digital systems in terms of transmission rate, we measure the number of voices that can be placed on a conventional telephone line. An analog system can handle just one voice. For digital systems, transmission at 9600 baud will be assumed. Where the data rate exceeds that, analog transmission from a synthesizer near the instructional computer giving one voice per phone line will be assumed.

The complexity of computation required to produce speech varies widely between synthesis techniques. Increased complexity can often increase the intelligibility and quality of the speech at an increase in expense.

Some factors directly affecting the cost of a system have been mentioned but producing accurate cost models of the various synthesis techniques is well beyond the scope of this paper. The cost is a most important criterion for comparison of CAI synthesis systems, but technology is changing rapidly enough to make evaluation on other criteria more important than on present day costs.

2.2 IMSSS Speech Requirements

Although any application would like the best possible speech, the current technology of speech synthesis does not allow for optimizing all the above criteria in a single system. Given this fact, different curriculums want to compromise in different ways. A foreign language course is more willing to give up manipulating intonation and composing messages for increased intelligibility, coverage and quality. Such a course would not want to expose its students to atypical pronunciations of the target language. However, a non-linguistic course using speech as a second source for stimulating and informing the student may well allow degradations in coverage, quality and intelligibility, if by so doing it is able to impart important semantic relationships through the intonation of its messages.

An evaluation of the curriculums at IMSSS (see Section 5 below) indicates that high quality and intelligibility are mandatory as is the ability to dynamically compose sentences and the random accessing of utterances. The ability to manipulate intonation is desirable. Since the mechanisms of changing intonation are not well understood, it is unrealistic to make this a mandatory requirement. Yet prosody manipulations are so desirable that importance is placed having some basic capabilities

---

(2) The Fairbank's Rhyme test and the modified rhyme test are suitable for testing intelligibility.[5]

(3) We use the words intonation and prosody interchangeably to refer to the speech's fundamental frequency (pitch), duration and volume contours.

to effect intonation. Since our curriculums have a fairly restricted vocabulary, it seems less important to choose a technique which allows word composition if such a method had substantially lower quality. Less important still were considerations of vocabulary storage size and data transmission rate since our students are all clustered near the computer and can share the same vocabulary storage device. Finally, the computational complexity of the synthesizer was not considered important due to the rapidly decreasing cost of computational hardware.

2.3 Description of Speech Synthesis Techniques

Many methods of reproducing and synthesizing speech have been used. Our division into five general techniques is primarily based on the methods' operational characteristics rather than their internal structure. The interested reader is referred to Flanagan's two books for detailed descriptions of the various synthesis techniques (see [6] and [7]).

The first speech system considered is a computer controlled analog magnetic tape player. It has excellent intelligibility, quality, and coverage, but has only linear accessing or at best locally random accessing. No sentence, word or intonation composition is possible. Tapes must be manually stored and retrieved. However, transmission of the control commands to the tape player requires an insignificant amount of a phone line's data capability and no computation is required.

Analog recordings on disks and drums would solve the tape's accessing limitations but due to the cost of such storage, most such systems have compromised quality, coverage and intelligibility for putting more messages on the available storage medium. Although sentence composition is possible, no ability to manipulate intonation is available.

Digital recording on disks and drums have exactly the same description in terms of our criteria of evaluation as their analog counterparts. Pulse Code Modulation (PCM), Delta Modulation, and Adaptive Delta Modulation are all popular digital representations of the analog speech signal, but none of these techniques are functionally different from an analog system except in the amount of magnetic material needed in the storage medium and the relative cost of implementing a particular system.

Acoustic modelling of the characteristics of speech signals provides a technique with significant improvements over simpler recording methods. Substantial reductions in storage size are accomplished by parametrically representing the short term speech spectrum and its voicing frequency. During recording, a sequence of approximately 50 such sets of parameters is recorded for each second of an utterance. For playback, synthesis is accomplished by using each set of parameters in turn to recreate a short period of the speech wave. Furthermore, as the synthesis is accomplished, variations in the speaking rate and voicing frequency can be accomplished without affecting the spectrum or phonetic quality of the speech. Thus, intonation can be manipulated.

Linear Prediction analysis and Format analysis are two popular methods of performing acoustic modelling (see [2], [8], [15], and [14]). Makhoul [13] has shown these two techniques to be very similar; however, Format analysis achieves a smaller storage size by assuming that the spectrum has a format structure. Lack of conformity to format structure results in lower quality speech than that produced using Linear Prediction. Furthermore, Linear Prediction analysis is computationally more direct than Format analysis. Good intelligibility, coverage and quality can be practically realized with either method since fewer compromises with storage size constraints are necessary. Digital disks and drums are natural mediums for storing the acoustic speech parameters. Sentences, words, and even parts of words can be recorded and stored as units. Message quality and intelligibility will be better with larger recorded units, but composability will be reduced.

Speech by rule forms words out of parts of words according to rules formulated for a particular language (see [4], [21]). Speech by rule makes use of the phonetician's concepts of phonemes and morphemes. Phonemes are classes of short speech segments where the class definitions are based on differences in the sounds which are recognizably distinct for the language. Morphemes are those combinations of several phonemes which provide minimal context for the interaction of phonemes (see [9] for a discussion of these last two points). Jon Allen's speech by rule system [1] synthesizes words by decomposing orthographic words into their morphological representation, then into phonological representation and finally into acoustic representation. To accomplish this, a lexicon of all common morphs and four sets of rules are needed. One set of rules is for decomposing the text representation of words into morphs, another set is for decomposing words directly into phones for those words which fail morpheme decomposition, a third set is for correcting the resulting phoneme sequence to take into account junctural effects, and a fourth set is for transforming the phoneme sequence into controls for a synthesizer. Speech by rule systems can have good intelligibility but due to rule failures, coverage is often not complete even for one language and coverage is certainly incomplete for multiple languages. Rule inadequacies cause quality to be fair. However, random accessing and dynamic composition of sentences, words and intonation are all possible. The vocabulary storage size is quite small in comparison with other synthesis methods, but a large dictionary of morphs is still needed. The transmission rate is low, but the computational complexity is very high.

2.4 IMSSS's Choice of a Speech Synthesizer

We decided to use a word based linear prediction synthesis system because of the following considerations: 1) the need for complete coverage by multiple language instruction prohibits using speech by rule systems; 2) the need for dynamic composition of messages by curriculums such

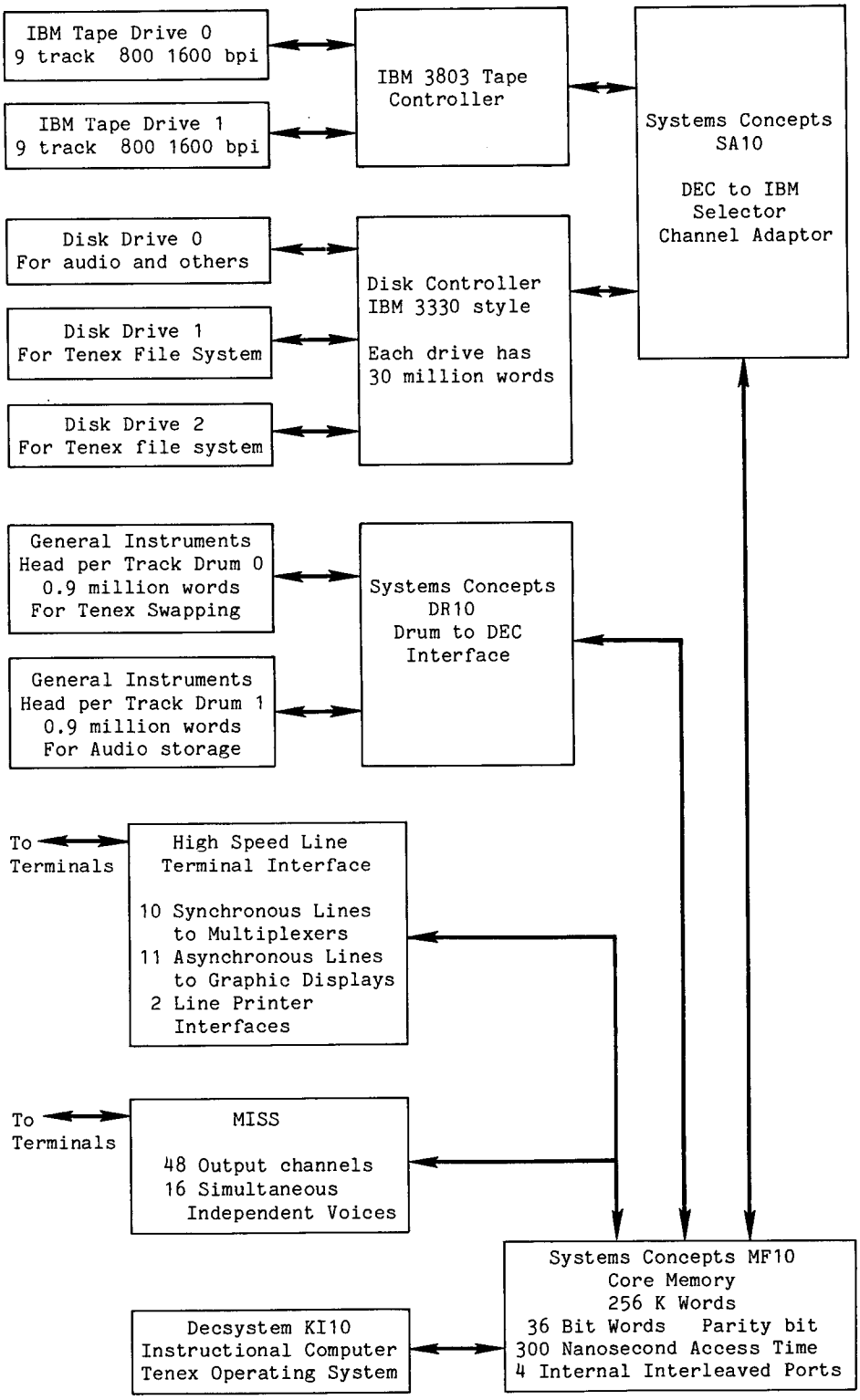

Figure 1. Block Diagram of the Instructional Computer System

as Set Theory prohibits using tape players; 3) the need to modify the intonation of a composed message prohibits using either analog or digital recording techniques. The choice of a word based system does place certain functional restrictions on applications. Intonation modifications are less precise and effective when applied only on word units instead of phoneme units. There is no ability to compose a word which is not in the recorded lexicon; however, some attempts will be made to combine affixes with root words.

Since no suitable synthesis system was commercially available, we implemented a linear prediction analysis program, designed and built a hardware synthesizer, and implemented software support suitable for use by curriculum programs.

3 Hardware Implementation of an Operational Synthesis System

A complete technical description of the hardware and software implementation is beyond the scope of this paper. Such a description will be available in a forthcoming technical report [17]. Here we present an overview of the implementation considerations important to the success of a CAI voice response system.

3.1 The Instructional Computer System

The instructional computer system at IMSSS consists of a central computer with its associated random access memory, a drum used for swapping, magnetic disks used for information storage, magnetic tapes used for off-line information storage, and user terminals used by both students and programmers. The computer is a Digital Equipment Corporation KI10 running the TENEX timesharing system. Core memory consists of a quarter of a million 36 bit words with an access time of 300 nanoseconds. The head per track drum provides an extension of memory to over 2 million words by using virtual memory and swapping techniques. On-line disk storage is over one hundred million words. Off-line storage uses conventional nine track tapes.

3.2 Hardware Additions Needed

The implementation of a word based synthesis system involved the addition of a second drum for vocabulary storage, the construction of a special purpose computer to perform the synthesis calculations, and the addition of amplifiers and headsets to the terminals. We decided that up to forty-eight terminals could benefit from speech capabilities. Based on our previous experience with a Delta Modulation speech system, we estimated that no more than sixteen of those terminals would require simultaneous speech. The head per track drum was needed because moving head disks are not fast enough to access the speech data for sixteen independent voices. The system with these additions is shown in Figure 1.

3.3 MISS Hardware

We call both the overall voice response system and actual synthesizer hardware MISS, which stands for Microprogrammed Intoned Speech Synthesizer. The synthesizer is actually a pair of very fast and closely coupled processors specially designed to efficiently perform the type of calculations needed. One of the processors is a microprogrammable digital filter which does the actual speech synthesis using linear predictive parameters. The output of the digital filter is converted to analog speech signals by any of forty-eight digital to analog converters whose outputs are transmitted to individual terminals. The other processor is a more general purpose computer which interprets commands from the instructional computer, retrieves speech data from the main memory, performs the appropriate intonation computations, and passes the data to the digital filter. A block diagram of MISS can be seen in Figure 2.

3.4 Terminal Hardware

Each terminal requiring speech has an amplifier and headset. The headset was chosen for its fidelity and isolation characteristics since several students often take lessons in the same room. The terminal headset amplifier is connected to the synthesizer by a pair of wires in much the same way as the terminal itself is connected to the computer's terminal multiplexer. In addition, synthesizer channels may be connected directly to phone line arrangements to service users via the telephone system.

4 Software Implementation of an Operational Synthesis System

We will briefly describe the vocabulary recording process, the support software written for use by the curriculum programs, the programs needed to experiment and develop the MISS system, and the data structures used in the operational system.

4.1 Vocabulary

Of primary importance to a word based synthesizer is the definition of the vocabulary. The English lexicon was composed from our previous audio system's vocabulary of 2700 words and was expanded to 12,000 words for other curriculums which had not previously used audio. A human speaker was chosen for the intelligibility and quality of his voice. He recorded the words directly onto the computer system using pulse code modulation at 240,000 bits per second to achieve a very accurate representation of the words. This recorded vocabulary was stored on digital tape. Later, the words were retrieved from tape, the Linear Predictive analysis was performed and the derived parametric representations were stored on both the drum and archival tape. This analysis reduced the storage requirements to 6000 bits per second of speech. Finally, the synthesized words are being verified to insure that each is typical of the intended word. Native Chinese and Russian

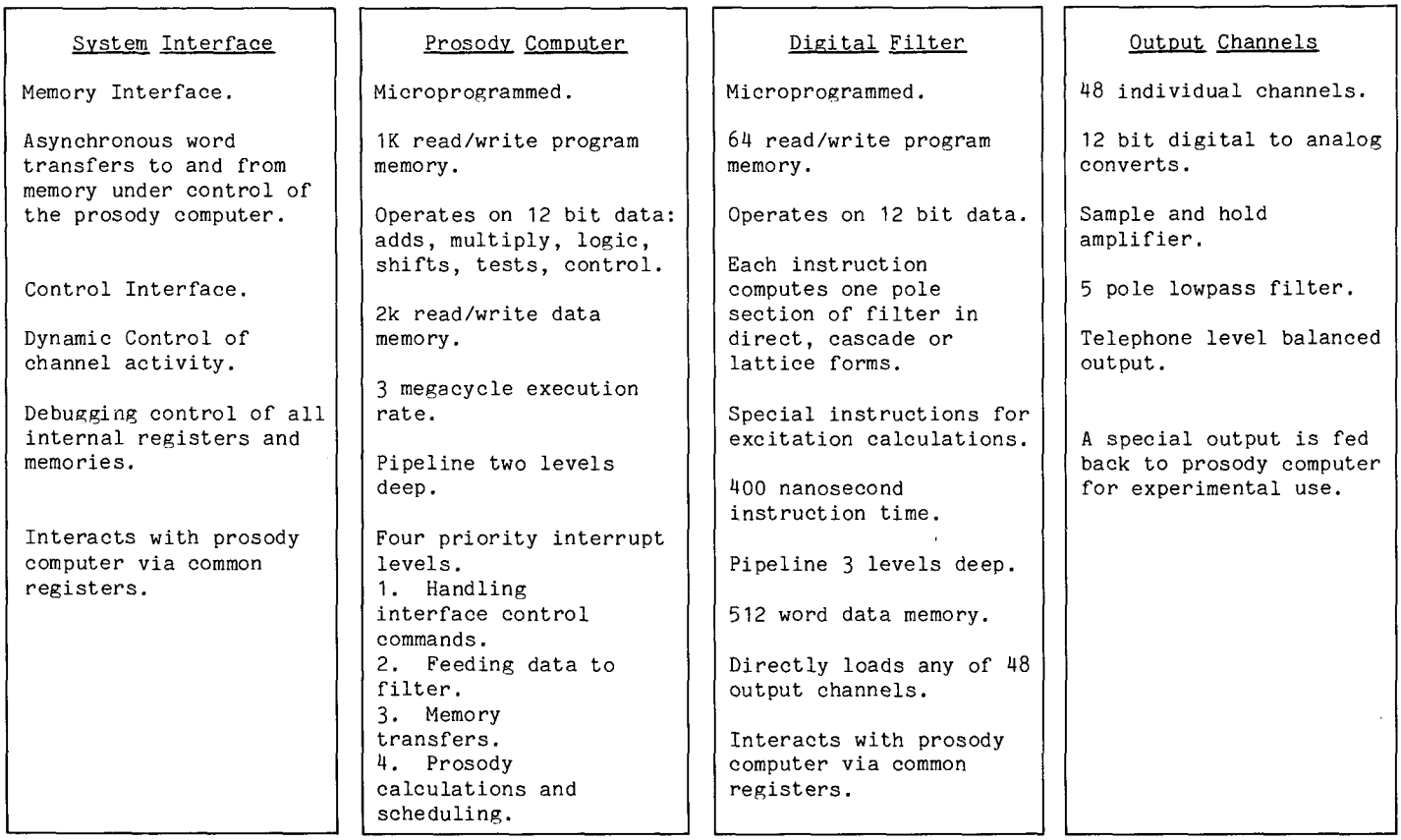

Figure 2. Block Diagram of the MISS Machine

Curriculum Program

The curriculum program composes a message and calls the library procedure SPEAK.
SPEAK("Every set has a power set.").

SPEAK Library Procedure

SPEAK breaks the argument into words. For each word it looks up the word's storage address in the lexicon and places them in a command list.
The command list:

1. address of EVERY
2. address of SET
3. address of HAS
4. address of A
5. address of POWER
6. address of SET

Lexicon

The lexicon is a hash table of all available words in a language. Entries in the table are addressed by name. Each entry has 11 fields which tell where the sound data is stored, the part of speech, and certain prosodic parameters of the word. This simple example does not show how the part of speech or the prosodic parameters are used to form prosody commands for MISS.

System Procedure

The system tells MISS where the command list is, then fetches the data from the drum/disk into a double buffer. As the data for each word is fetched, the corresponding command list entry is changed to reflect where in memory the data is. MISS uses this change to know it is possible to start retrieving the data. When MISS is finished with a buffer, it zeros out the command list entry; this tells the system the buffer is free and available for fetching the next sound.

System Buffers

The command list and buffers are shown as the MISS machine is saying "HAS" and the data for "A" is being fetched from the drum.

Command list	Buffer 1	Buffer 2
0 (completed EVERY)	Data for the	This buffer
0 (completed SET)	word HAS.	is being
pointer to buffer 1	MISS is reading	filled by the
address of A	out of this	drum with the
address of POWER	buffer.	data for the
address of SET		word A.

Figure 3. Flow Diagram of Speech Data and Commands

speakers have recorded small vocabularies for use in language courses. These vocabularies will be analyzed like the English vocabulary.

4.2 MISS Software

Software to access and manipulate the words was written as programmed library procedures. These procedures were carefully documented so all instructional programs using audio could access the speech in a uniform and efficient way. The use of library procedures has facilitated making changes in the speech data and synthesizer, since the effects of these changes to the various CAI programs are usually hidden inside the procedures.

Curriculum programs only have to specify the language to use and the level of automatic prosodic manipulations to be performed. Whenever the program wants a particular message spoken, it simply passes the text representation of the message to a library procedure. This procedure is described in the next section.

Acoustic programs are interested in more sophisticated manipulations of the speech. They want to record, edit, move, analyze, and transform the speech data. They also need to edit the lexicon, and graphically display the various representations of speech. Furthermore, these programs want to be able to run both in interactive mode for experimental use and in batch mode for automatically processing large vocabularies. The library procedures provide mechanisms for all these operations in such a way that programs may easily incorporate these features.

We have three main acoustic programs, one for recording and analyzing the vocabulary, one for experimenting with intonation, and one for manipulating and archiving lexicons. The analysis program takes a word, finds where the speech begins and ends, detects the precise locations of all pitch pulses in the speech, performs the Linear Predictive analysis and does further data reductions on these parameters. The analyzed versions of the words are then stored on the drum, the disk, or tape and the text representation of the word is stored in a lexicon. The intonation program allows for manipulating the pitch, duration and volume of sequences of words to study their effect on the perception of sentence prosody. These manipulations will eventually be performed by the MISS processor. The lexicon program is used for defining homonyms, defining special expanded pronunciations of words, and archiving sounds onto magnetic tape.

4.3 Data and Command Structures

A sound may be represented in terms of its Pulse Code Modulation values, its Delta Modulation values, or its Linear Predictor Coefficents. The sound data is stored in a uniform format on all storage devices. The sound data itself is stored, along with additional information naming the actual speech representation used and the methods used to transform the speech into this representation.

The lexicon provides a method of addressing by name data on the disk, drum, or tape. Each language has its own lexicon, and each lexicon is associated with a particular region of a storage device. The desired name can be efficiently looked up in the lexicon, and information associated with the name referenced. Each name has associated with it up to ten information fields, three of which are used to specify the storage device and address of the sound data. Other fields are optional and can contain linguistic information needed for parsing and applying intonation contours, or other information needed by a particular application. Lexicons are implemented as shared program segments under TENEX, so all programs using a language share the same physical memory.

The efficient flow of commands and data through the computer system is essential for an operational system. To achieve efficiency, a system must minimize data movements and computations. Since an overly efficient implementation can inhibit the system's flexibility and extensibility, a careful balance must be maintained. Figure 3 shows the flow of information needed for a typical message. A program that dynamically composes a message to be heard at a student terminal passes the text representation of the message to a library procedure. This procedure converts the text into a command list consisting of the drum or disk addresses of the speech data for each word interspersed with intonation commands. The system causes the data to be retrieved from the storage device and placed in a memory buffer. The memory location of the data and the intonation commands are passed to the synthesizer. MISS retrieves the commands and sound data as it needs them, applies the intonation commands to the data, and passes the resulting data to the digital filter. The filter performs its computations which generate a Pulse Code Modulation version of the speech. This is passed to an output channel which converts this data into analog speech. The speech is transmitted to a student wearing a headset at a terminal. Command lists can be prepared by a preprocessor for messages that are not created dynamically, such as standard greetings and error messages, thus improving efficiency.

5 Applications

In Section 2 we discussed the requirements that we placed on an audio system. We stressed there that our overall need is to compose intelligible messages from randomly accessed words, with sufficient quality for foreign language instruction.

The CAI systems currently under development at IMSSS emphasize dialog with the student concerning a semantic structure built up interactively. We refer to this approach to the use of computers in education as the use of complex instructional systems (see [24]). For example, the EXCHECK system teaches set theory by building a high-level model of the student's proof.

5.1 Previous Uses of Audio in Computer-assisted Instruction

Several of the methods of producing sound discussed in Section 2 have previously been used instructionally, and are still being used. We will not give a complete history of this, but will mention a few examples.

Analog tape has been used frequently. Audio tape was also used for a number of years in the teaching of Russian at Stanford. A bank of six high-quality tape machines were put under computer-control. Tapes keyed to the curriculum were mounted on the tape decks by an operator. This system gave quality reproduction, but the messages were fixed in both order and content, with the curriculum designed to move in lock-step fashion. This system, without basic changes, was later implemented with cassette tape hardware.

Analog tape was also used in the Stanford-Brentwood Computer-assisted Instruction Laboratory [23], a project that combined film, audio, and CRT output under computer control. There were 17 instructional stations, and a computer-controlled tape drive associated with each station. Lessons in the mathematics curriculum covered such topics as counting, numerals, linear measure, and so on.

Audio messages were coordinated to the CRT display. For example, in one lesson in mathematics, a car and a truck surrounded by set braces appeared on the screen, accompanied by the audio "There are two members in this set." After the message, two more sets appeared on the screen, one empty and the other containing a train and a steamshovel. The student heard the instructions "Find another set with two members". To answer correctly, the student moved his light pen to the area of the screen with the correct set, and the computer responded "Yes, the sets have the same number of members". (See pp. 271-301 of [23] for more details here.) Messages were contingently programmed to depend on the student's response to the problems. The tape drives offered locally random accessing and good quality, but no composition.

Specially designed tape devices have been put to varying instructional uses. For example, Ishida and Fujimura [10] used a hybrid analog-digital device for instruction in foreign language pronunciation. The system provided locally random accessing.

Analog disks and drums have also been used, perhaps less frequently than tapes. For example, a drum-type system was used at IMSSS to teach spelling (see [12]). The particular system, a Westinghouse Prodac-50 controlling 12 drives with six-inch wide tapes, allowed a degree of random access. The quality of sound was good but the equipment required inordinate maintenance and was hence abandoned. A disk system was implemented for use with the PLATO IV system. Fixed messages (for several curriculums) could be randomly accessed. The system has been used experimentally in teaching elementary reading and veterinary medicine. This system was considered too unreliable, but a new version of the system is planned. [19]

Digital PCM audio was used in a system called Dial-a-Drill, a product of the Computer Curriculum Corporation, as reported in [11]. New York City students were called at home and given daily drills in elementary mathematics over the telephone. The students responded to the audio messages by typing the keys of a touch-tone telephone.

Delta modulation was used very successfully at Stanford in teaching elementary reading (see [3]). The reading curriculum was divided into strands, each of which was "designed to provide practice on a particular decoding or communication skill" ([3], p. 5). Typical student dialog consisted of a Teletype terminal printing out a list of words

BIKE LIKE STRIKE

while the audio message told the student to "TYPE STRIKE". Suitable audio and printed reinforcement was given when the student typed the correct answer.

The main instructional innovation of the reading program was the use of a complex statistical model that optimized the student's instruction. This helped produce the "significant and consistent gains in reading achievement over what would be expected from classroom instruction alone" ([3], p. 1).

The use of audio in reading was fundamental to the program. Over 2700 words and short phrases were available to the program, which randomly accessed those messages. The Delta system served well, but the quality was poor and the coverage was incomplete. We should also point out that student dialog was minimal. The messages were either just a word or a short phrase, and the program did not create a semantic data base of any sort about which to have a dialog with the student.

The Delta system was also used for teaching French at Stanford. The curriculum structure was based on eight strands for conceptual categories, such as gender, pronunciation, and idioms. Within each strand, students received drill on patterns, such as:

AUDIO: Paris est une ville.

AUDIO: un lion ... un animal

In the above, the student used the first message as a pattern, and the second message as the words for that pattern. The correct response from the student would have been for him to type:

UN LION EST UN ANIMAL.

The inability to compose messages (other than by concatenating the sounds) made creating patterns like the above quite difficult. The results obtained were achieved only by re-recording the sounds of the individual words until some of the patterns fit together reasonably well. Furthermore, words recorded by the same speaker on different days did not combine as well as words recorded at the same time. The lack of prosodic control extends well beyond this, however, into the semantics and pragmatics of the message.

Raugh, Atkinson, and Schupbach [16] used the Delta audio system in controlled experiments to determine the effectiveness of the keyword method for learning foreign language vocabulary. As part of the learning method, the computer spoke the words of the target languages (Spanish and Russian in different experiments), establishing an acoustic link in the student's memory to an English keyword. The Delta system was efficient enough to serve in this kind of application, which did not require composed messages.

5.2 Using Audio in Language Instruction

MISS will be used for instruction in Mandarin Chinese in a course developed by Mr. E-Shi (Peter) Wu. This course is now being experimentally used with the old Delta Modulation audio system, and will be used with MISS in early 1976. This work is briefly described in [24]. The conversion to MISS will enable us to experiment with using intonation manipulations to create the different tones of Chinese.

We will be able to evaluate the importance of the higher quality and more complete coverage of MISS by using it with curriculums which had previously used the Delta Modulation system. We expect that MISS's better speech will, in itself, only slightly improve these curriculums. The real benefit of MISS will be the ability to say things that previously could not be said.

5.3 Contributions of Speech to Program-Student Interaction

We have identified three separate models of program-student interaction using computer controlled speech as a main component. Each of these models arises from a real instructional problem in the logic and set theory courses at IMSSS. These three models are:

> 1) The computer gives directions to the student about how to use the computer in some way. Here, the audio component takes the place of a human instructor who might sit beside the student guiding him through an initial dialog.
>
> 2) The computer gives an explanation, using graphic display and audio simultaneously.
>
> 3) The computer and the student are in full dialog, with some appropriate parts of the conversation using the audio component.

We now give some details of the three audio projects underway with regard to logic and set theory courses at Stanford.

5.4 Using Audio to Give Directions

A common problem for the student is getting started using an instructional program. It is difficult to remedy this using only the student's terminal as the medium of communication, since the assistance gets confused with the object dialog. Common techniques to handle this kind of instruction include the use of manuals and human assistance. We have experienced this problem with the set theory program, for example. The manual provides sample dialogs and complete command specifications, but we have generally made personal appointments with the students on various occasions throughout the course, particularly at the beginning, in order to familiarize them with the operation of the set theory program.

We are designing a program OVERVIEW that will solve this problem in part. It operates by putting the object program (in this case, the set theory program) into a job structure separate from the OVERVIEW program. Thus it can observe the operation of the object program, generating audio messages about the dialog. It is important that OVERVIEW operate somewhat asynchronously from the object program, and that the object program not have to be extensively modified for this purpose. The disadvantage of OVERVIEW is that it will have only limited access to the internal components of the object program. (Some of the data structure will be shared, through features of TENEX.)

5.5 Using Audio in Proof Explication

Another problem we have noted in set theory is the difficulty the student experiences in keeping a clear notion of the proof he is constructing, especially when the argument becomes lengthy. We have developed a procedure that takes a lengthy proof, finds the "important" steps of that proof, and produces a sketch or summary. There are many proof-theoretic and linguistic problems in obtaining this analysis, which are discussed elsewhere.

When the student requests a sketch of his proof, the program will produce a graphic display (see [20] for a sample of such a display). This display shows the main steps of the proof (or subproof) and their relationship. The audio will then further summarize what is on the screen, paraphrasing formulas and giving derivation chains informally.

For example, consider the problem of paraphrasing sentences of set theory. If the (somewhat formal) sentence

> For every x there is a y such that
> y = powerset of x

is a part of the derivation, we can paraphrase this with the audio message

> Every set has a powerset.

The basis of this particular paraphrase is 1) dropping of explicit variables not linguistically necessary, and 2) using the verb "to have" to indicate function application. We know how to generate for output purposes many informal paraphrases that we do not know how to recognize when input.

The audio, in this application, will give

explanations that are quite informal. For example, consider the formula:

```
For all m,n,p
        m times n = p
           If and only if
        There are A,B such that
            (Card(A) = m & Card(B) = n
                AND
             Card(A X B) = p)
```

A reasonable audio message for describing this formula might be:

```
Defines cardinal multiplication
                    using cross product.
```

It is reasonable because it might provide just the intuition that the student needs to understand the formula. Teachers in the classroom tend to write in precise detail on the blackboard while giving a less precise intuition verbally. The use of audio here is analogous to this explanation role of the teacher.

This proof explanation system is being developed and should be operational in early 1976.

5.6 Using Audio in Full Dialog

IMSSS has for a number of years had CAI programs for elementary logic. The current program uses some of the advanced proof machinery from the set theory program in order to make the instruction more informal and at a logically higher level. (See [24] for a short history and some details of the current program.)

A deficiency of the current curriculum is the lack of instructional material about semantics. A common teaching method in logic texts and courses is to give the semantics of first-order logic in natural language paraphrases and examples. In traditional courses, the student is often called upon to formalize sentences and arguments starting with an English formulation.

Inspection of the fragment of English involved in several texts (for example, [22]) indicates that it is sufficiently uniform for us to handle using our current methods of natural language processing. We are producing a dialog system in which students can formalize sentences and arguments and then deal with these arguments in English and formal logic.

In doing this, it is important to keep the "object language" of first-order logic separate from the "meta-language" of the English sentences. This ability will be provided by the audio system, which will inform the student of the intended interpretation of the formulas and proofs he is constructing. Paraphrases, similar to those described above in the discussion of set theory, will be used.

This application of audio will be ready for students in late 1976.

5.7 Summary of Applications

The above initial applications of the audio system are all directed to current problems in the logic and set theory courses. All of them share the following properties:

1) The vocabulary is relatively small and fixed.

2) The dialog will be based on the constructions that the student is doing, not some fixed set of messages in a branching network. This is particularly true in the latter two examples.

3) It is very important to be able to alter sentence stress patterns. We believe the ability to paraphrase formulas and summarize proofs depends on control of stress.

References

1. Allen, J. Speech synthesis from unrestricted text. In J. L. Flanagan & L. R. Rabiner (Eds.), Speech Synthesis, Stroudsburg, Pa.: Dowden, Hutchinson, & Ross, 1973.

2. Atal, B. S. & Hanauer, S. L. Speech analysis and synthesis by linear prediction of the speech wave. Journal of the Acoustical Society of America, 1970, 50, 637-655.

3. Atkinson, R. C., Fletcher, J. D., Lindsay, E. J., Campbell, J. O., and Barr, A. Computer-assisted instruction in initial reading (Tech. Rep. No. 207). Stanford, Calif.: Institute for Mathematical Studies in the Social Sciences, Stanford University, 1973.

4. Coker, C. H., Umeda, N. & Browman, C. P. Automatic synthesis from ordinary English text. IEEE Transactions on Audio and E acoustics, June 1973, AU-21, No. 3, 293-298.

5. Fairbanks, G. Voice and articulation drillbook (2nd ed.). New York: Harper & Brothers, 1940.

6. Flanagan, J. L. Speech analysis synthesis and perception. New York: Springer-Verlag, 1972.

7. Flanagan, J. L. & Rabiner, L. R. (ed.), Speech synthesis. Stroudsburg, Pa.: Dowden, Hutchinson & Ross, 1973.

8. Gray, A. H. & Markel, J. D. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. IEEE Transactions on Acoustics, Speech and Signal Processing, June 1974, ASSP-22, No. 3, 207-217.

9. Heffner, R-M. S. General phonetics. Madison, Wisc.: The University of Wisconsin Press, 1969.

10. Ishida, H., & Fujimura, O., A computer-based pronunciation-hearing test system using a hybrid magnetic tape unit, Proceedings of the IFIP Congress 1971, V. 2, Amsterdam: North Holland, 1971.

11. Jerman, M., Clinton, J. P. M., & Sobers, A. W. A CAI program for the home. Educational Technology, December, 1971, p. 49.

12. Lorton, Paul Jr., Computer-based instruction in spelling: An investigation of optimal strategies for presenting instructional material. Ph. D. dissertation, Stanford University, June, 1973.

13. Makhoul, J. I. Natural communication with computers: Speech compression research at BBN. Bolt Beranek and Newman, Inc., December 1974, Final Report, V. II, Report No. 2976.

14. Makhoul, J. Spectral analysis of speech by linear prediction. IEEE Transactions on Audio and Electroacoustics, June 1973, AU-21, No. 3, 140-148.

15. Markel, J. D. & Gray, A. H. A linear prediction vocoder simulation based upon the autocorrelation method. IEEE Transactions on Acoustics, Speech, and Signal Processing, April 1974, ASSP-22, No. 2, 124-134.

16. Raugh, M. R., Schupbach, R. D., & Atkinson, R. C. Teaching a large Russian vocabulary by the mnemonic keyword method (Tech. Rep. No. 256). Stanford, Calif.: Institute for Mathematical Studies in the Social Sciences, Stanford University, 1975.

17. Sanders, W. R. & Benbassat, G. V. The MISS speech synthesis system, in preparation. Institute for Mathematical Studies in the Social Sciences, Stanford University.

18. Smith, R. L., Graves, H., Blaine, L. H., & Marinov, V. G. Computer-assisted axiomatic mathematics: Informal rigor. In O. Lecarme & R. Lewis (Eds.), Computers in education, part 1: IFIP. Amsterdam: North Holland, 1975.

19. J. Risken, Personal communication, CERL, University of Illinois, Champagne-Urbana, Illinois.

20. Smith, R. L. & Blaine, L. H., A generalized system for mathematics instruction, in press.

21. Stevens, K. N., Kasowski, S. & Fant, C. G. M. An electrical analog of the vocal tract. In J. L. Flanagan, & L. R. Rabiner, (Eds.), Speech synthesis. Stroudsburg, Pa.: Dowden, Hutchinson & Ross, 1973.

22. Suppes, P., Introduction to logic, New York: Van Nostrand, 1957.

23. Suppes, P., and Morningstar, M., Computer-asssisted instruction at Stanford, 1966-68: Data, models, and evaluation of the arithmetic programs, New York: Academic Press, 1972.

24. Suppes, P., Smith, R. L., & Beard, M. University-level CAI at Stanford: 1975 (Tech. Rep. No. 265). Stanford, Calif.: Institute for Mathematical Studies in the Social Sciences, 1975.