[redacted]

SPERRY UNIVAC SPEECH COMMUNICATIONS TECHNOLOGY

MARK F. MEDRESS

SPERRY UNIVAC DEFENSE SYSTEMS DIVISION
ST. PAUL, MINNESOTA

INTRODUCTION

During the past nine years, the Speech Communications Research Department at Sperry Univac has been developing technology and systems for effective verbal communication with computers. The department has nine professionals trained in the speech sciences, linguistics, and computer science. A versatile laboratory computer facility is dedicated to speech research activities, and is complemented by a large and powerful time sharing system. Major projects include the development of a continuous speech recognition system for verbal input, a word spotting system to locate key words in conversational speech, prosodic tools to aid speech analysis, and a prerecorded voice response system for speech output. The primary focus of this paper is on our speech recognition system. Brief descriptions of our other speech projects, as well as our resources for speech technology development, are also included.

CONTINUOUS SPEECH RECOGNITION

A primary goal of our speech research has been the development of a linguistically oriented computer system for recognizing naturally spoken phrases and sentences[1-4]. In contrast to currently available isolated word recognizers, our system does not require users to either pause artificially between words, or to repeat every vocabulary word several times for system training. It is also able to recognize speech from a number of similar talkers without adjustments for individual voice characteristics. With suitable vocabulary and syntactic restrictions, the recognition of a wide variety of connected word sequences for practical speech input applications will be possible in the near future. Because of the linguistic framework used for recognition, the system can gracefully evolve to understand more natural sentences with the enhancement of syntactic and semantic analysis capabilities.

The Recognition System

The principle components of the speech recognition system being developed at Sperry Univac are shown in Figure 1. In the first step of the recognition process, the speech waveform is digitized with a 5 kHz bandwidth, and an acoustic analysis is performed with autocorrelation,

[redacted]

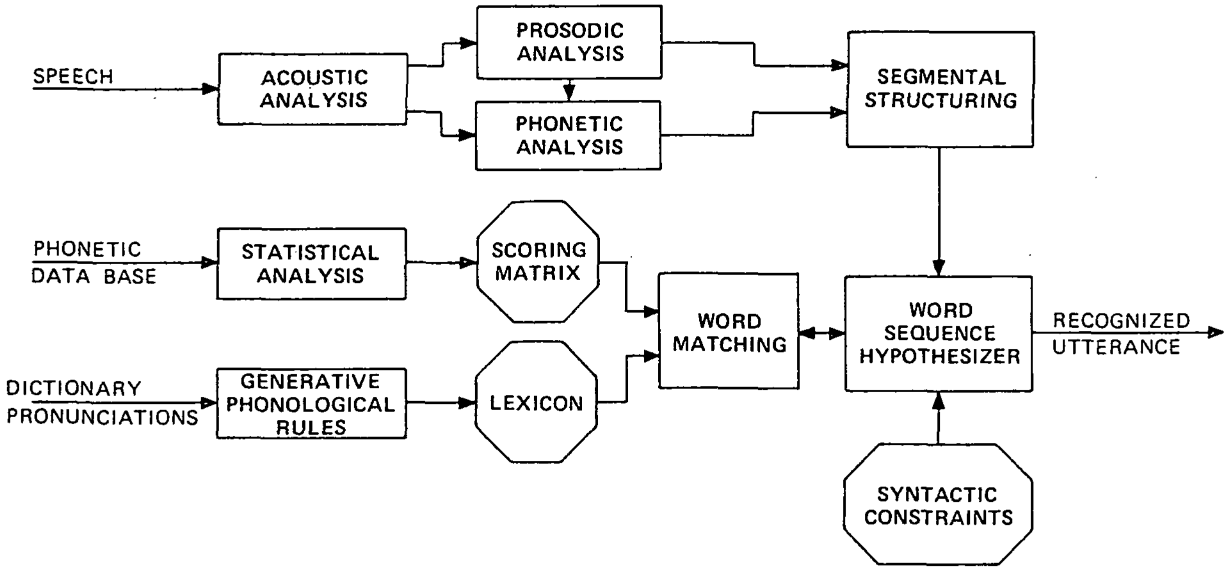

Figure 1. The Sperry Univac Continuous Speech Recognition System

Fast Fourier Transform, and linear prediction processes to produce 14 time functions that describe voice fundamental frequency, bandlimited energies, and vocal tract resonances, or formants. Next, a prosodic analysis component provides information about the syllabic structure of the utterance, including the preliminary locations of syllabic nuclei, as well as estimates of which syllables are stressed. A phonetic analysis component then determines the sound segments, or phonetic sequences, throughout the unknown utterance, including the locations and subclassifications of stops, sibilants, nasals, vowels, liquids, glides, and fricatives[5]. This phonetic feature information is represented in a two-dimensional lattice of sound classes versus time. In preparation for vocabulary matching, a segmental structuring component next transforms the lattice of phonetic information into a non-overlapping sequence of analysis segments, making various phonological or segmental adjustments during the transformation.

To complete the recognition process, a word sequence hypothesizer determines which sequence of vocabulary words best matches the analysis segments of the unknown utterance[6]. It uses syntactic constraints to direct a word matching component, which aligns and scores segments from each word in the dictionary, or lexicon, with the appropriate analysis segments. The lexicon itself is produced by a generative phonological rules component, which automatically transforms standard dictionary pronunciations into likely alternative sequences of analysis segments[7]. Using vowels as anchor points and allowing both missed and extra segments with appropriate penalties, the word matcher aligns and scores the analysis and lexical segments with the aid of a scoring matrix, which is generated by a statistical analysis processor that correlates analysis segments with time-locked phonetic transcriptions for a data base of development utterances. Working from left to right, the word sequence hypothesizer then strings together good single word matches. The best scoring sequence of words that spans all the analysis segments and satisfies the syntactic constraints, is chosen as the recognized utterance. (A more detailed description of this system can be found in Reference 4.)

A Recognition Example

Figure 2 illustrates how the phrase "six seven nine" is recognized by our system. After acoustic, prosodic and phonetic analysis, the segmental structuring component produces the twelve analysis segments shown at the top of the figure. The analysis vowels, which serve as anchor points for lexical matching, are enclosed in solid boxes. Beginning with the first analysis vowel, the word sequence hypothesizer directs the word matcher to find and score all syntactically permitted lexical matches, allowing for missed, extra, and incorrectly identified segments. High scoring matches are then extended by anchoring around subsequent vowels, until the best scoring sequence of lexical entries is found. Note that in hypothesizing word sequences, the matcher accommodates continuous speech by specifically allowing consecutive words that end and begin with similar consonants, to share consonantal analysis segments.

In this example, the lexical entry for "six" (enclosed by a dashed box) is aligned around the first vowel as shown. The alignment is scored by computing the average of the segment scores, which are given in the figure between the analysis and lexical segments. Each score is the logarithm of the estimated conditional probability that the particular lexical segment was spoken, given that the corresponding analysis segment was found. To extend the sequences beginning with "six," the lexical entries are next aligned with the second analysis vowel, and the result for "seven" is illustrated. The word sequence hypothesis beginning with "six seven" is completed by aligning lexical entries with

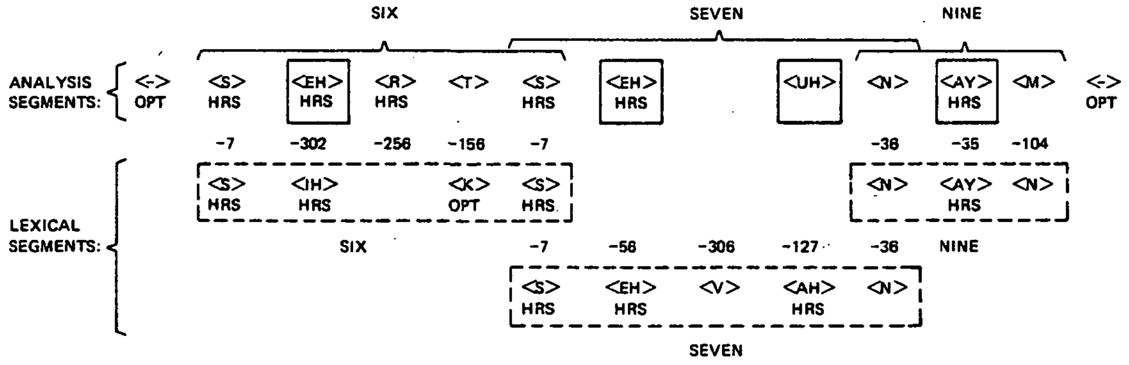

Figure 2. Recognition of the Phrase "Six Seven Nine"

the fourth and final analysis vowel, as the result for "nine" shows. While many alternative word sequence hypotheses are considered, the best scoring sequence for this example is that presented in Figure 2, and the utterance is therefore correctly recognized.

The Recognition Data Base

During the past year, our continuous speech recognition system was developed and tested on a speech data base representing two application areas. The first of the task domains consists of two, three, and four word sequences of digits and "phonetic alphabet" words, a vocabulary and syntax characteristic of many data entry tasks. The 36 word vocabulary is divided into four subsets of eight to ten words, and nine varieties of sequences are defined. Examples of these "alphanumeric" sequences are listed in Figure 3. The average branching factor (average number of word alternatives to the right of each word of the sentence) for this task is 9.4. The syntax defines 25,842 potential sequences.

The second task addresses the recognition of utterances typical of data management or information retrieval languages, and is based upon a potential speech input application in air traffic control. The seven "command" types listed in Figure 3 define the permissable syntactic structures. The items in parentheses are fixed one-word subsets for that utterance type, while the underlined words are variable subsets consisting of the numbers 1-9, 10-19, or 20-90 by tens; the positions "up", "down",

ALPHANUMERIC SEQUENCES

- Vocabulary size = 36
- Average branching factor = 9.4

e.g. Hotel niner
Sierra Alfa Zulu
Quebec Papa four three

DATA MANAGEMENT COMMANDS

- Vocabulary size = 64
- Average branching factor = 6.3

1. (Shift line) twelve (to) (position number) ten
2. (Transmit line) eighteen (to) (station) two
3. (Cursor) down seven
4. (Erase) field
5. (Flight index for) American forty nine
6. (Weather forecast for) Minneapolis
7. (Current weather for) Boston

Figure 3. Sample Phrases and Sentences for Speech Recognition Development and Testing

"left", or "right", the objects "field", "line", or "page"; ten airline names; and ten city names. The total vocabulary size is 64, and the average branching factor is 6.3. The syntax defines a potential of 919 different utterances.

For each task domain, 111 utterances were randomly selected for recording and processing. Three male talkers each recorded about one-third of the utterances. Approximately two-thirds of the data base was used for developing the recognition programs, and the remaining third was reserved as test material. No adjustments of the recognition system were made for individual talker characteristics.

Recognition Performance and Future Development

After the development system was stabilized, the test data portion was processed to obtain test results. For both the alphanumeric sequences and the data management commands, the results are shown in Figure 4 for the correct recognition of the individual words in each phrase, as well as for the correct recognition of the complete phrases. The number of words and phrases in each category is given in parentheses beside the percentage results. For the alphanumeric sequences, the correct phrase recognition was 91% for the 75 development phrases and 83% for the 36 test phrases. For the data management commands, the correct phrase recognition was 95% for the 74 development phrases and 78% for the 37 test phrases. The overall results are 88% correct for the alphanumeric sequences and 89% correct for the data management commands.

Within the next few years, we expect to improve our speech recognition system so that it can meet the performance requirements of a variety of practical applications for continuous speech input. Our current recognition system operates in about 300 times real time on our laboratory minicomputer, with approximately 95% of that time devoted to acoustic analysis. The system should operate in real time with the planned addition of a fast array processor, and with more efficient use of our minicomputer's hardware and software capabilities. Recognition accuracy should also increase as the result of incorporating both phonetic analysis

ALPHANUMERIC SEQUENCES

Speech Data	% Correct Individual Word Recognition	% Correct Phrase Recognition
Development	97% (225)	91% (75)
Test	93% (108)	83% (36)
Average	95% (333)	88% (111)

DATA MANAGEMENT COMMANDS

Speech Data	% Correct Individual Word Recognition	% Correct Phrase Recognition
Development	98% (256)	95% (74)
Test	91% (128)	78% (37)
Average	96% (384)	89% (111)

Figure 4. Word and Phrase Recognition Performance for the Development and Test Sentences

improvements based on context information, and a word verification component being developed under another project. Studies already under way of noisy, bandlimited speech should eventually lead to successful recognition over telephones and other communication channels. All of these planned improvements are designed to provide an effective and practical sentence recognition system for natural speech input to computers.

OTHER SPEECH COMMUNICATIONS PROJECTS

In addition to its development of a linguistically oriented continuous speech recognition system, Sperry Univac has been involved in several related and complementary speech development activities. These include projects for word spotting, prosodic research, and voice response.

Word Spotting

Our word spotting project is a major research activity that is using many of the same components and technologies from our continuous speech recognition system to develop procedures for spotting key information-carrying words in natural conversations[8]. While the simple location of selected words is a more limited task than that of recognizing all the words in a conversation, several new attributes make this a challenging problem indeed. First, the talker population is large, unknown, and non-cooperative; it includes both men and women with a wide variety of dialects and acoustic characteristics. Second, the speech is very informal and conversational, and is therefore characterized by large fluctuations in amplitude, speaking rate, and articulatory preciseness. Finally, the conversations are conducted over normal telephone channels, so the resultant speech has limited bandwidth, added noise, and other spectral and temporal distortions imposed by the communication medium.

A block diagram of our word spotting system is shown in Figure 5. The similarity between this system and the one we are developing for continuous speech recognition should be apparent from a comparison of Figures 5 and 1. The acoustic analysis, prosodic analysis, phonetic analysis and segmental structuring components produce a linear sequence of analysis segments representing the conversational speech material. While these components are basically the same as the corresponding ones in our speech recognition system, they are being suitably modified to better handle the limited signal bandwidth and wide variety of talkers[9]. The word hypothesizer is also similar to that of our other system. Again using vowels as anchor points, it aligns and scores keyword representations from a segmental lexicon with the analysis segments, to determine where in the incoming speech are likely occurrences of keywords. Each hypothesized keyword occurrence is then further evaluated by a new component developed for our word spotting system. Using dynamic programming for time registration, this word verifier provides an independent assessment of the acoustic

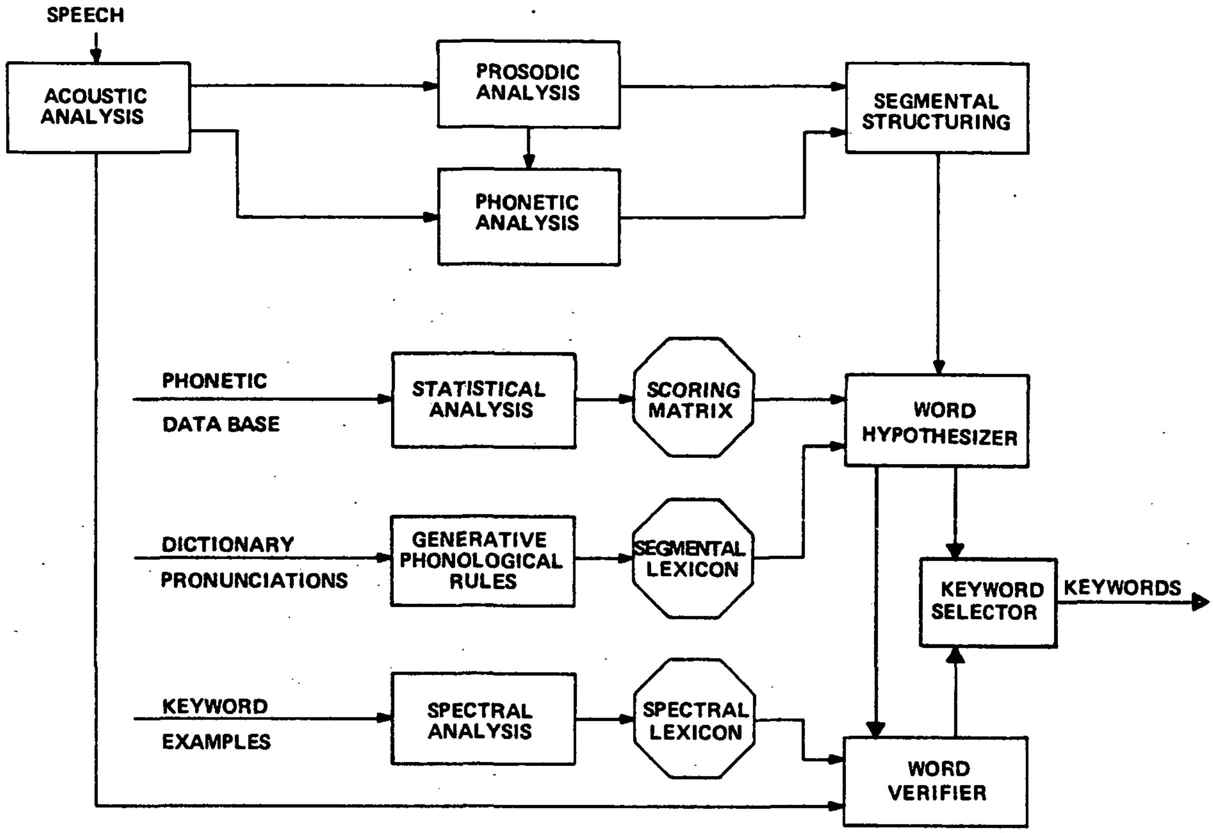

Figure 5. The Sperry Univac Word Spotting System

similarity of a stored spectral pattern for the hypothesized word, with the spectral characteristics of the input speech at the region hypothesized. A novel feature of our verifier is its use of vowel nuclei for anchoring the alignment process. Finally, a keyword selector operates on the word scores provided by both the hypothesizer and verifier to produce a list of accepted keywords and their locations. (Reference 8 contains a more complete description of this system.)

An initial version of our word spotting system has been developed on 13 minutes of informal telephone conversations by eight talkers, and tested on 11 additional minutes of speech by two of the same talkers and eight new ones. Results of this test are encouraging, and development is continuing with a focus on improving acoustic and phonetic processing and word verification. The current test materials will be folded in as new development data, and the system will be retested using speech from 16 additional talkers. Studies are also under way to extend the system so it can perform acceptably with noisier speech.

Prosodic Research

Besides its continuous speech recognition and word spotting development activities, Sperry Univac has also participated in a five-year Speech Understanding Systems Program funded by the Advanced Research Projects Agency (ARPA) of the Department of Defense[10-11]. Our research in this project centered on the development of prosodic aids to speech recognition and understanding systems[12]. We formulated procedures for using such prosodic information as intonation patterns, stressed syllable locations, and speech rhythm in a speech understanding system for natural sentences[2]. Programs were developed to segment continuous speech into major syntactic phrases based on fall-rise valleys in voice fundamental frequency contours, to locate syllabic nuclei in regions of high energy bounded by substantial dips, and to associate syllabic stress with those high-energy syllabic nuclei near the initial fundamental frequency rise in each phrase, and near substantial fundamental frequency inflections at later points in the phrase. Some of these programs have been incorporated into our own speech recognition and word spotting systems, as the block diagrams in Figures 1 and 5 indicate. Studies were also conducted of how such prosodic information could be used in other speech understanding systems developed in the ARPA program, especially the system at Bolt Beranek and Newman.

Voice Response

The projects described so far have all centered on the computer analysis of speech, with a major application being for verbal input to computers. Sperry Univac's voice response developments address the opposite problem: the computer generation of high quality, natural sounding sentences for speech output. Instead of creating speech by synthesis methods, our prerecorded voice response units use words and phrases that are first spoken by a trained announcer and then digitized and stored in a digital memory, as shown in Figure 6. To produce speech output, a host computer first specifies the sequence of words and phrases that form the desired output message. The voice response controller next retrieves the digitized speech from the vocabulary storage memory and strings the

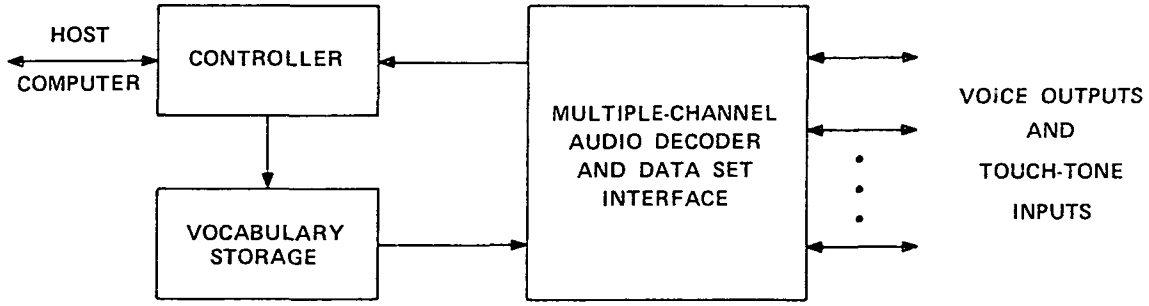
HOST
COMPUTER
CONTROLLER
VOCABULARY STORAGE
MULTIPLE-CHANNEL AUDIO DECODER AND DATA SET INTERFACE
VOICE OUTPUTS AND TOUCH-TONE INPUTS

Figure 6. The Sperry Univac Voice Response Unit

specified words and phrases together without undesireable intervening pauses. The audio decoder and data set interface portion then converts the digitized speech back into an analog signal, and the resulting voice output message is sent to a speaker, radio transmitter, or telephone circuit. The voice response unit is also able to accept touch-tone input characters for internal use or for transmission back to the host computer.

Our latest voice response unit, the VRU-400, is completely solid state and has several attractive features[13]. The controller is implemented with a programmable microprocessor, providing a great deal of flexibility and internal processing capability. The vocabulary is stored in a solid state memory made of Charge Coupled Device (CCD) memory chips, resulting in increased reliability, faster access, and better modularity than a disk-based unit. By using Adaptive Differential Pulse Code Modulation (ADPCM)[14], we are able to obtain high quality digitization of telephone bandwidth speech using only 24 kilobits per second of vocabulary, about half the bit rate needed with ordinary PCM encoding. The speech output quality is further enhanced by using variable-length vocabulary storage, and by composing messages from complete phrases whenever possible. We also record two versions of some vocabulary items, one version with flat inflection for use in the middle of a phrase, and the other with falling inflection for phrase-final position. The basic VRU-400 can handle up to 16 simultaneous and independent audio-output/touch-tone-input channels, and a vocabulary of up to 200 seconds of recorded speech. Additional vocabulary can be accommodated with extra vocabulary storage memory.

A number of practical applications have been successfully addressed by Sperry Univac's voice response units. They have been used by the Federal Aviation Administration to automatically generate voice messages in their air traffic control systems[15]. Typical examples include traffic advisories, metering and spacing messages, and minimum safe altitude warnings. The National Weather Service and the Department of Transportation have also used our voice response units to provide pilots with information about current and predicted weather conditions. Finally, we have recently installed a VRU-400 in a telephone ordering system for a large catalogue retailer in the Federal Republic of Germany. The voice response unit allows customers to place their orders over ordinary telephones, using touch-tone signals for input, and voice response messages (in German) for output. The voice response unit, which is on-line to the main order-processing computer, provides real time confirmation of the item ordered, its availability, and its current price. Merchandise delivery time has also been significantly reduced since the VRU-400 eliminates mail delays in placing orders.

RESOURCES FOR SPEECH TECHNOLOGY DEVELOPMENT

As a result of Sperry Univac's growing involvement in a variety of speech projects over the past nine years, we now have substantial resources available for developing speech communications technology. These include competent and experienced personnel, and excellent computer and laboratory facilities.

Personnel

The present staff of the Speech Communications Research Department consists of nine professionals with a variety of relevant backgrounds in acoustics, phonetics, phonology, syntax, semantics, system design, and hardware implementation. Dr. Mark Medress, Dr. Timothy Diller, Dean Kloker, and Toby Skinner all have graduate training and a great deal of experience in speech science and linguistics. Don Anderson and Dave Andersen are experienced system design engineers who have been responsible for our voice response projects. Laboratory and software development support are provided by Henry Oredson, Larry Lutton, and John Siebenand. Together the department members have over 65 years of cumulative and productive involvement in speech and natural language processing.

Facilities

The Speech Communication Research Department has over 3,500 square feet of office and laboratory space in Univac Park, the headquarters of Sperry Univac's Defense Systems Division in St. Paul, Minnesota. Complete laboratory facilities are available for speech research activities, including a sound isolation room for a controlled audio environment, a Voicescan spectral analyzer for making speech spectrograms, a versatile dedicated minicomputer system, and terminals connected to a large and powerful time sharing system. Most of the laboratory facilities are contained in a special environment that provides the highest level of physical and electromagnetic security, thereby permitting both unclassified and classified projects to be properly accommodated.

A block diagram of our dedicated minicomputer system, called our Speech Research Facility (SRF), is shown in Figure 7. It consists of a Sperry Univac 16-bit minicomputer, a Hardware Fast Fourier Transform processor (HFFT), normal peripherals for program development and storage, and an interactive control console and graphic display, in addition to modules needed to support Sperry Univac voice response systems that are deployed in the field. With the SRF, speech can be digitized and stored, converted back to audio and played over a speaker, and displayed on a CRT. Spectra, time functions, and other parameteric results obtained from the speech waveform can also be viewed on the graphic display, as can intermediate and final results of speech recognition programs. Full interactive control of the SRF is provided by a large number of push-

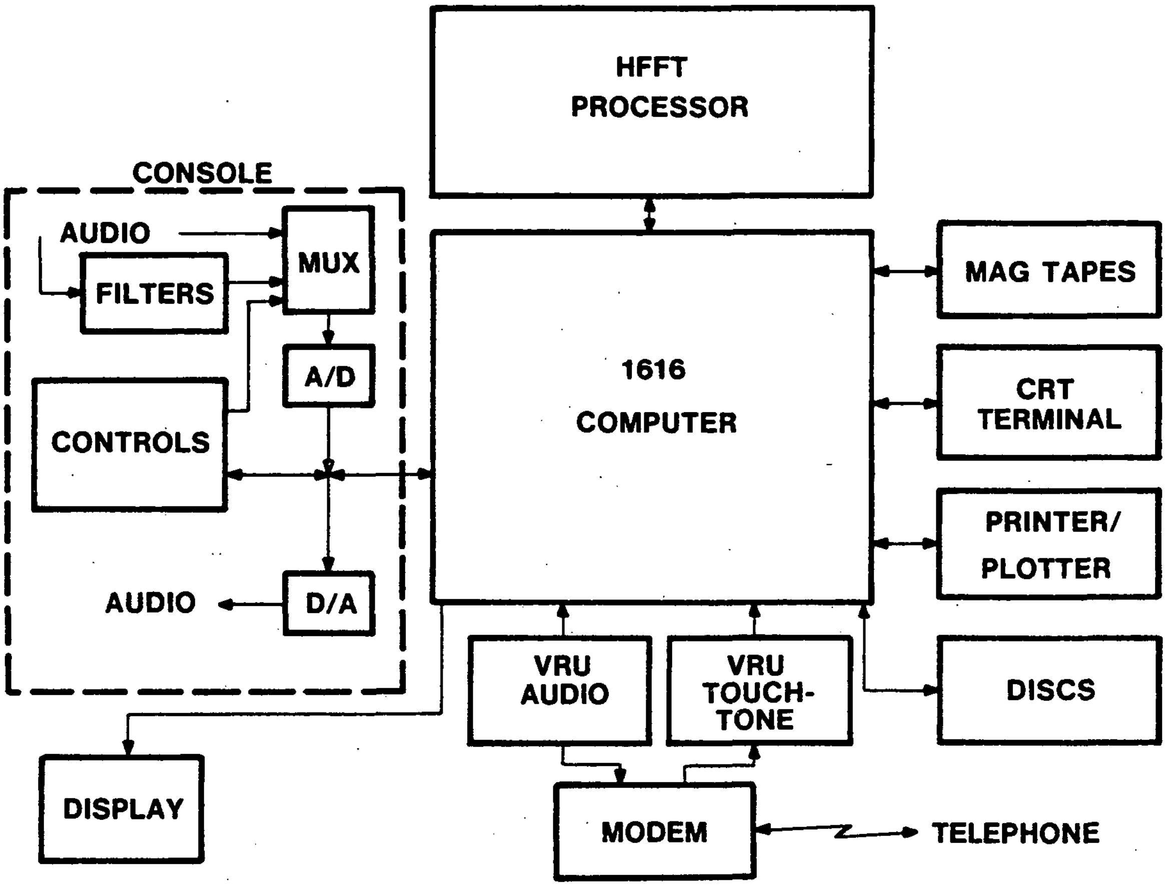

Figure 7. The Sperry Univac Speech Research Facility

buttons and potentiometers, as well as an alphanumeric keyboard and display. Analog filters provide bandlimited energy functions in real time, and together with the HFFT, permit fast and efficient complex processing of speech.

In addition to the SRF, a functionally equivalent software system (without A/D, D/A, and interactive graphics capabilities) has been implemented on a time-shared Sperry Univac 1100/43 computer facility. The Speech Communications Research Department has six terminals connected to this facility, a large amount of disk file storage, and effective procedures for transferring programs and data between the 1100 and our laboratory minicomputer. This time sharing capability allows multiple users to develop and test algorithms and procedures, and to choose the most effective computer system for each task.

SUMMARY

Sperry Univac is developing technology that will make computer systems easier and more natural to use, by providing them with effective verbal input and output capabilities. A continuous speech recognition system is under development for understanding naturally spoken phrases and sentences by a number of talkers. Current recognition performance is very encouraging, and we expect a practical version of this system to be available for a variety of continuous speech input applications within a few years. Another major project is developing a related system for locating key information-carrying words in natural conversations by a large and diverse group of people communicating over standard telephone lines. High quality, natural sounding speech output is already available with our VRU-400, a solid state voice response unit that has been successfully tested in air traffic control, weather broadcasting, and telephone ordering applications. Our past accomplishments, as well as our potential for future progress in developing speech communications technology, are a result of both a well trained and experienced staff, and excellent research facilities. And since Sperry Univac's Defense Systems Division is a major supplier of ruggedized computer systems to the Department of Defense and other government agencies, we are able to effectively integrate emerging speech technology into these systems, thus bridging the gap between the research laboratory and practical applications in operational environments.

REFERENCES

1. Medress, M. F. (1972). "A Procedure for the Machine Recognition of Speech," Conference Record of the 1972 IEEE Conference on Speech Communication and Processing, IEEE Cat. No. 72 CHO596-7 AE, pp. 113-116.

2. Lea, W. A., Medress, M. F., and Skinner, T. E. (1975). "A Prosodically-Guided Speech Understanding Strategy," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-23, pp. 30-38.

3. Skinner, T. E., Kloker, D. R., and Medress, M. F. (1976). "A Speech Recognition System for Connected Word Sequences," Conference Record of the 1976 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE Cat. No. 76 CH1067-8 ASSP, pp. 434-437.

4. Medress, M. F., Skinner, T. E., Kloker, D. R., Diller, T. C., and Lea, W. A. (1977). "A System for the Recognition of Spoken Connected Word Sequences," Conference Record of the 1977 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE Cat. No. 77 CH1197-3, ASSP, pp. 468-473.

5. Skinner, T. E. (1977a). "Toward Automatic Determination of the Sounds Comprising Spoken Words and Sentences," Sperry Univac Report No. PX 12124.

6. Kloker, D. R. (1976). "A Connected Word Sequence Matching Strategy for Speech Recognition," Sperry Univac Report No. PX 11649.

7. Diller, T. C. (1977). "Automatic Lexical Generation for Speech Recognition," Conference Record of the 1977 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE Cat. No. 77 CH1197-3, ASSP, pp. 803-806.

8. Medress, M. F., Diller, T. C., Kloker, D. R., Lutton, L. L., Oredson, H. N., and Skinner, T. E. (1978). "An Automatic Word Spotting System for Conversational Speech," Paper presented at the 1978 IEEE International Conference on Acoustics, Speech, and Signal Processing.

9. Skinner, T. E. (1977b). "Speaker Invariant Characterizations of Vowels, Liquids, and Glides Using Relative Formant Frequencies," Paper presented at the 94th Meeting of the Acoustical Society of America.

10. Medress, M. F., Cooper, F. S., Forgie, J. W., Green, C. C., Klatt, D. H., O'Malley, M. H., Neuburg, E. P., Newell, A., Reddy, D. R., Ritea, B., Shoup-Hummel, J. E., Walker, D. E., and Woods, W. A. (1977). "Speech Understanding Systems," IEEE Transactions on Professional Communication, Vol. PC-20, pp. 221-225.

11. Klatt, D. H. (1977). "Review of the ARPA Speech Understanding Program," Journal of the Acoustical Society of America, Vol. 62, pp. 1345-1366.

12. Lea, W. A. (1976). "Prosodic Aids to Speech Recognition: IX. Acoustic-Prosodic Patterns in Selected English Phrase Structures," Sperry Univac Report No. PX 11963.

13. Anderson, D. E., and Andersen, D. P. (1977). "The VRU-400/MP Voice Response Unit," Sperry Univac Report No. PX 12270.

14. Anderson, D. E. (1977). "ADPCM-Coded Speech for Voice Response Systems," Sperry Univac Report No. PX 12181.

15. Beck, A. F., and Anderson, D. E. (1975). "Computer-Generated Voice in Air Traffic Control Applications," Proceedings of the IEEE 1975 National Aerospace and Electronics Conference NAECON '75, IEEE Cat. No. 75 CHO956-3 NAECON 75, pp. 547-551.

BIOGRAPHICAL SKETCH

Mark F. Medress

Mark F. Medress obtained his B.S., M.S., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology in 1965, 1968, and 1969, respectively. His doctoral thesis, under the direction of Professor Kenneth N. Stevens, involved the development of a phonetically-based word recognition system. After completion his graduate studies, he joined Sperry Univac to participate in speech research and development activities, and has been Manager of the Speech Communications Research Department since 1972. Dr. Medress was also a member of the steering committee that coordinated the Speech Understanding Systems Program of the DOD Advanced Research Projects Agency, and served as acting committee chairman toward the end of the program.

DISCUSSION

Dr. Mark Medress

Q: Rex Dixon, IBM: What is the data rate of ADCPM coding that you're using?

A: We're running on about 24 kilobits. We're sampling 6,000 samples per second using four bits per sample.

Q: Don Connolly, FAA: What kind of processing times are you talking about on these connected sequences?

A: Good point and I forgot to mention it. The version that we had running last spring was 300 times real time on this mini computer that I showed you in the block diagram. It means that if you set a two second utterance, it took 600 seconds to complete the recognition. We have a version of that system almost integrated that will run about 150 times real time; and on this mini computer system, I think our limit is about 20 or 30 times real time. But we'll also be buying a processor that will do our acoustic analysis in real time and that's 95% of our processing. It will also be useful in doing word verification and some of our signal matching searching procedures.

Q: Steve Moreland, Army Aviation R&D Command: You mentioned that you were recording messages for this voice response system. I would like to hear a little more explanation. You're not doing synthesized voice but you're doing something else, right?

A: Right. We're doing pre-recorded voice. Every word or phrase that has to be strung together to make a sentence has to be first spoken by a person, put on an analog tape, digitized, and stored away in a vocabulary memory.

Q: Steve Moreland: Then you're not calling up a recorder to play back or anything of that nature. You're actually in essence synthesizing it, aren't you?

A: No. It's just like a computer control tape recorder but its digital with random access. I'll be glad to explain it to you in more detail.

Q: Steve Moreland: O.K. Have you measured the speech intelligibility from that?

A: No, we haven't but we've gotten very good reaction to it from people who have either heard it or used it in their applications. It's very high quality. I've got a tape that I'd be glad to play for you later if you like.

Q: George Doddington, TI: Rather than change the subject, let me ask a question about speech synthesis. Apparently, from what you said about LSI, CCD, storage and whatnot, storage is a problem. So why not do a synthesis from a very low bit rate data rather than say 24 kilobits?

A: We probably will within the next year. The reason that we stuck with the ADCPM at this point is because we wanted a short term, easy to implement and high quality system. I should say that our customers wanted that. There are people here who are much more highly versed and experienced in aero bound speech representations or compressed speech representations than I am so that is a very relevant question and we're interested in doing that in fact. We're interested in replacing ADCPM with linear prediction analysis synthesis or something like it to reduce storage requirements.

Q: Jared Wolf: In your word spotting, word verifier component, how do you derive those word storage spectral templets?

A: The spectral templets come by excising examples of key words that we're looking for from actual occurrences in the development data base; and in fact, what we did was we took all the occurrences of the key words in the development data and correlated them against one another; that is all the tokens of a particular word which is correlated by the word verifier using dynamic programming and so on. To find which ones matched each other well and where there were different subsets, in the 10-word lexicon, we actually have 12 patterns. We have eight words that are represented by one pattern each and two words that are represented by two patterns each. And this is for a data base of 16 talkers including males and females.

Q: Leon Ferber: That means at one point, you couldn't have two false alarms? That means that one key word excludes all others.

A: No, I didn't talk about it at all but in fact for each vowel in the analysis segments we look for all possible words from the dictionary. We're looking for 10 in fact. So we test each of the 10 words against the area around that vowel and for each word there is a threshold of acceptability and each word that we've tested that exceeds its threshold is reported as a key word. So the one vowel might be 10 key words.

Q: George Doddington: O.K. Now that we're back on speech recognition, let me ask you the question. I assume you're working on the performance of improving your speech recognition technology so in that context I would like to know what your opinion is about what is the weak link? What are you working on?

A: O.K. That's a good question. I'll try to answer it with two responses. One is we really are interested in improving our acoustic phonetics analysis capability. And this fairly consistent with what Wayne Lea said has been reported to him from the ARPA program and from what you and I have talked about in the past. I think we do a pretty good job of acoustic phonetic analysis but we would like to do a better job. We feel for the very constrained sentence type that we're dealing with our matching capability is really fairly good but we would like to do a better job of the analysis phase, the phase or system that produces segments. And the other thing that we're very anxious to do is to incorporate our word verifier. Because one of the problems with the phonetic word analysis procedure is that you're throwing away information and you have to deal with co-articulation in order to do a good job of analyzing the segments and if you propose a word and can go back and verify that proposal by looking at the details in the spectrum throughout that word you can hopefully do a better job of saying this is a good hypothesis or this isn't a good hypothesis. So those are kind of the two major areas.

Q: George Doddington: Well, what about segmentation? I thought you were going to say segmentation is a difficult problem.

A: Oh, I'm sorry. That's what I meant by acoustic phonetic analysis. The process of getting a string of segments that represents input speech. What I call the analysis segments in the description of our system. I can show you in more detail later.

C-2