THE ARMY WORD RECOGNITION SYSTEM

DAVID R. HADDEN
DAVID HARATZ

U.S. ARMY COMMUNICATIONS RESEARCH &
DEVELOPMENT COMMAND (PROVISIONAL)
FORT MONMOUTH, NEW JERSEY

1. INTRODUCTION

In many cases, the spoken word still represents man's most effective and efficient means of communication. Either in the input or output mode, the spoken word can represent an important methodology for man/machine interactive communications in future army tactical communications, command and control (C^3) systems as we move into future weapons and control systems heavily dominated by digital computer technologies. The major areas of applicability for word recognition systems within the field army encompass aspects relevant to overall system man/machine reliability as compared to other interactive senses, i.e., touch, sight, etc. Although previous efforts have been primarily aimed at providing two-way verbal communications between front-line personnel (i.e., artillery forward observer) and an army tactical data system, near-term future directions will attempt to primarily tackle a subset of this very complex problem in terms of shelter oriented terminals encompassing console operations applicable to a wide range of source data automation (SDA) problems.

The Army's initial effort in the application of speech technology to Army tactical data systems was called Word Recognition System (WRS). The WRS was addressing the problem of the frontline troops (artillery forward observer) who must get information to a computer based system (i.e., TACFIRE). The primary thrust in this area was the keyboard entry/display device called the Digital Message Device (DMD). The alternative of the WRS was very appealing as it meant that the forward observer need only carry the radio as in a non-automated system and the required processing could be colocated with the action computer. The requirement for the WRS was therefore a limited vocabulary in a structured message format but with a variety of individual voices and Army tactical communications.

II. PRESENT STATUS

An advanced development model of a Word Recognition System (WRS) was developed for the Army aimed at providing two-way verbal communications

[redacted]

between front-line personnel and Army Tactical Data Systems (ARTADS) using discrete word recognition, speaker identification/verification and voice response techniques. The minicomputer based WBS is capable of fully automated real-time prompting, message translation and synthesized-speech response over a communication net for any of 64 users with a vocabulary of approximately 150 words. The vocabulary and syntax are a subset of messages that are capable of simulating forward observer inputs to TACFIRE via a Digital Message Device (DMD).

The WRS recognizer is an acoustic pattern classifier that produces a digital code as an output in response to the received utterance. It consists of a spectrum analyzer, an analog multiplexer and A/D converter (ADC), a programmed digital processor, a reference-pattern memory, and an output register as shown in Figure 1. The spectrum analyzer divides the input audio spectrum into 16 frequency bands that cover the useful frequency range. By means of parallel detection and lowpass filtering the resulting 16 analog signals represent a power spectrum that constitutes the feature for speech classification. These 16 continuous signals are multiplexed, sampled at 100 Hz, and converted to digital form with 8-bit precision. Thus, the original utterance arrives at the digital processor as a string of 8-bit binary numbers. The coding compressor compensates for changes in the rate of articulation and reduces the spectral data generated by each utterance to a fixed-length code for the classifier. It reduces every word, regardless of length, to a 240-bit pattern; for a word lasting two seconds the data compression exceeds 100 to 1. As a result, the fixed-length codes can be processed in real time by pattern-recognition techniques. The compression algorithm is essentially an arithmetic process which preserves selected property changes during an utterance and eliminates periods during which these properties remain constant.

The word boundary detector serves to establish the start and end of each utterance for the compressor by means of experimentally determined criteria. During the training or adaptation phase of system operation, approximately five utterances of each vocabulary word are elicited from the user. The estimator compensates for variations among these utterances to form a single, 120-bit reference pattern, and a 120-bit mask that is stored in memory to represent a particular vocabulary word. These 240 bits represent both the tendencies that are common to the five utterances and the small variations that are inevitable from utterance to utterance.

After the system has been trained to a particular user, each new 120-bit pattern from the coding compressor is compared with a syntactically determined subset of all the previously learned reference patterns in memory. Basically the classification process matches the patterns bit by bit via an exclusive-OR function that produces an output for each matching pair of bits. The total number of outputs is counted for each of the reference patterns; the one pattern that produces the

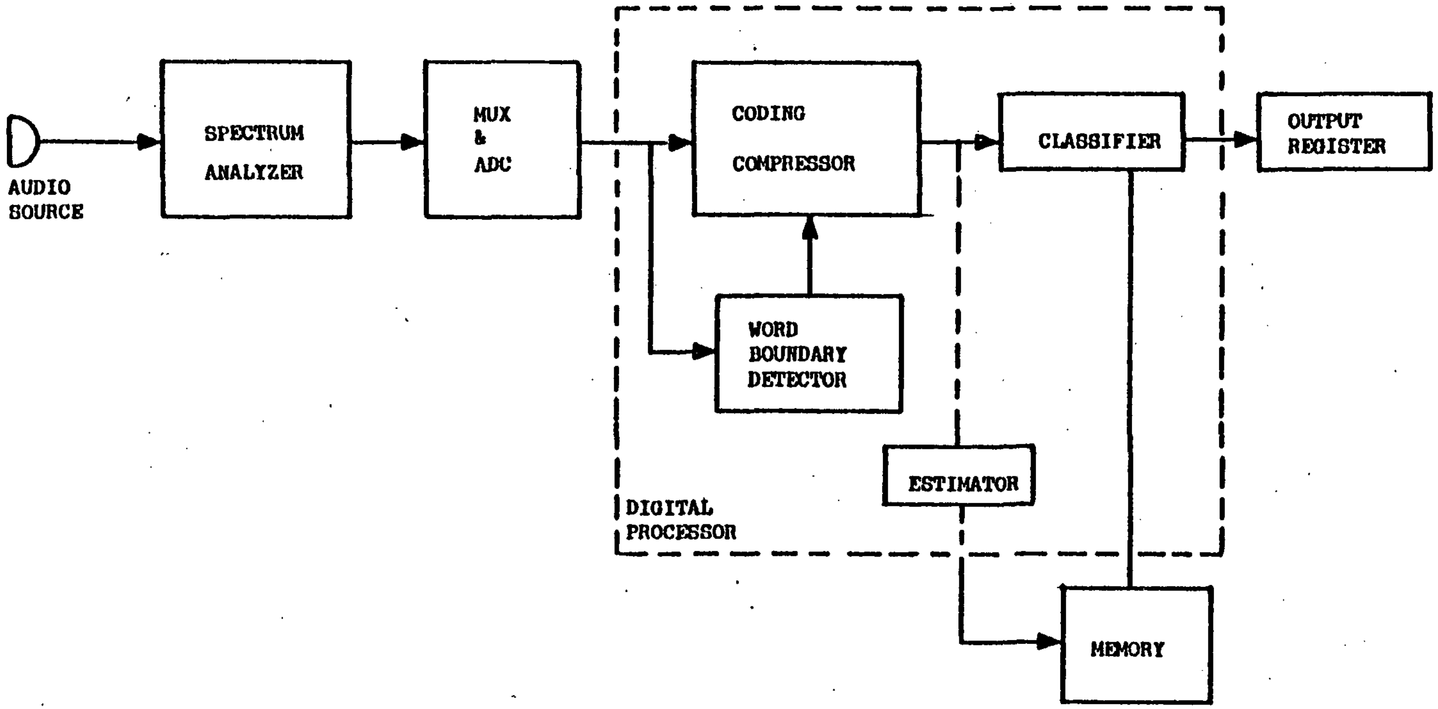

Figure 1. Word Recognizer in Army WRS

greatest number of outputs above a preset reject threshold classifies the compressor output and thereby the received utterance. The voice-response technology in WRS is an off-the-shelf Votrax module from the Vocal Interface Division of Federal Screw Works, Troy, Michigan. It is an audio synthesizer programmed to produce strings of phonemes and phoneme-like sounds with inflection to produce words.

Training or adaptation of the system to each of the 64 users may take place at the WRS site or via the communications link. The entire recognition vocabulary of about 150 words must be trained by each user by means of approximately five repetitions of each word. This vocabulary is divided into groups of about 30 words each to ease the training regimen. A single word retrain capability also exists. The user is prompted throughout the training regimen by voice response from the system.

A. Hardware Design

In the area of hardware design, the driving factor was the requirement for a field-deployable, ruggedized system to demonstrate the overall feasibility of word recognition in a semi-tactical environment. This necessitated the selection of an existing ruggedized processor and memory, the Rolm 1602. The hardware consists of the 1602 processor and memory, a number of standard peripherals, and the WRS preprocessor that was designed and fabricated by the contractor, SCOPE Electronics Inc., Reston, Virginia.

The CPU and its piggyback memory chassis occupy one drawer. An external memory chassis occupies another. Main memory is 32K 16-bit words of core, which is expandable to 64K as required through the addition of another piggyback memory chassis behind the Rolm 2143. The moving-head disk controller and disk drive are driven from the processor chassis; the disk is utilized to store user reference patterns and the operational software. A third Rolm Chassis provides access to the I/O bus for most of the peripherals. The CRT terminal serves as the operator display and control device. The line printer is the principal output device used to simulate messages transmitted to ARTADS. The magnetic tape unit is a backup storage and loading device for user reference patterns and operational software. The card reader, the paper tape reader, and the backup ASR33 terminal are used solely for software development.

All non-digital interfaces are made to the WRS preprocessor, which occupies one drawer and includes all of the front-end hardware for speech recognition and the computer interface speech synthesis. The hardware configuration of the WRS preprocessor and associated Votrax modules is shown in Figure 2. Each of the three net inputs (only one is implemented) may interface with: 1) a field telephone, 2) either of two types of non-secure FM radio transceivers, or 3) either of two types of secured FM radio transceivers. The net control module provides the necessary R/T switching functions and permits the operator to monitor any one, none, or all three nets at this option. A connection is brought to the

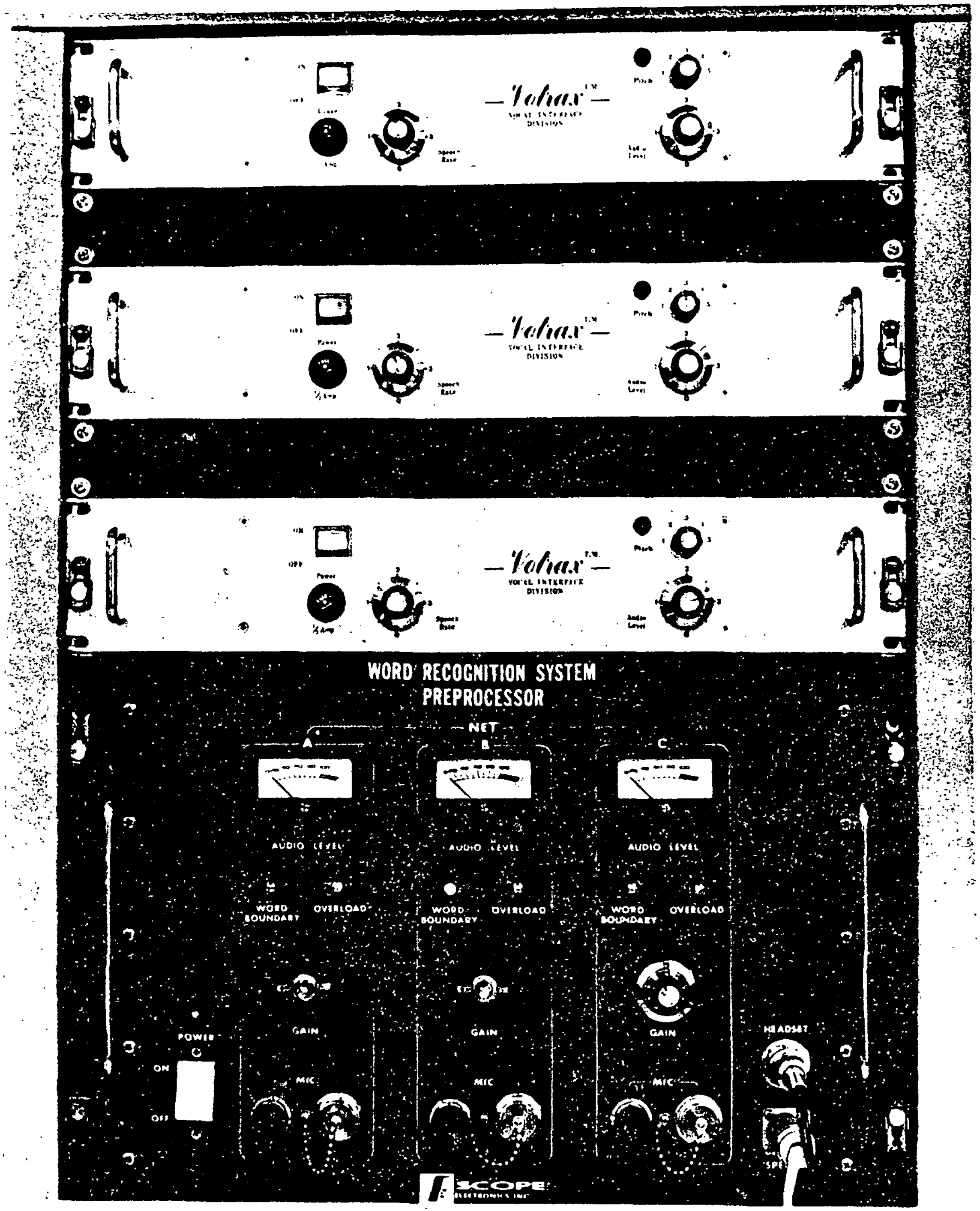

Figure 2. WRS Preprocessor Module and Votrax

front panel for the operator headset and a demonstration speaker that reproduces sound in the operator headset. A foot switch is provided for the operator's push-to-talk function. A voice response unit interface module was developed for communication between the I/O bus and the three voice response units (one per net).

Input speech signals arrive at the audio modules from either the communications nets or test microphones. Each audio module amplifies and modifies the incoming frequency spectrum to compensate for communications degradation and user microphone characteristics. Each of the three filter/ ADC modules then performs the spectral analysis, MUX, and ADC functions as previously described. A single ADC interface module controls the sampling, channel selection, and A/D conversion processes by software command.

Two modular power supplies provide the requlated DC power required by the preprocessor. A frequency-selectable source for the real-time clock permits experimental variation of the speech sampling interval. Preprocessor front panel indicators include (for each net) a meter, a word-boundary indicator that illuminates during each word, and an overload indicator to reveal saturation of the A/D converter.

B. Software Design

In the area of software design and implementation, the two primary requirments imposed encompassed the necessity for user independence and for system flexibility. The term "user independence" means that the performance of the system, as seen by one user, is unaffected by the status of the other system users. In this definition, the operator is considered as the fourth user of the system, since he is in competition with the three potential "verbal" users for the system facilities and must also be serviced in a manner which does not degrade with increased usage of the speech channels.

Because the purpose of WRS was to provide a system which was to be used to determine the suitability of this approach to field use, a high premium was placed on overall system flexibility. This requirement was addressed by a design philosophy which required that all application-dependent parameters be input from the outside--via a system configuration tape--rather than from the inside (that is, imbedded in the program itself). Thus, although the system was configured to operate within the constraints of the TACFIRE (and TOS[2]) syntax structures, it may easily be adapted to changes in these specifications or to entirely new vocabulary and syntax structures.

The executive system provides for both user independence and system flexibility through a structure based on the use of a separate user-state array for each user. This array contains all the variables needed by the system to completely describe the operating state of the user and also includes locations which may be used by each user to store application-dependent data during speech processing. Because all pertinent variables are located in a continguous array, the addition of some simple commands allows the system to manipulate the user-state array and, therefore, the user's actual status in a very flexible manner.

This flexibility may be enhanced even more if the commands which are used to manipulate the user-state array can be input as data and triggered through a variety of inputs. Because different applications of the executive system will involve potentially different vocabulary sizes, syntax structures, and even numbers of users, it is important that the allocation of memory be as flexible as possible. The executive accomplishes this goal through use of a set of memory management routines that allocate memory on an as-needed basis. Thus, any required memory tradeoffs may be made during system configuration.

In addition to providing the facilities described above, the executive system also supports two file structures that are used to describe the syntax and vocabulary for a particular application. One file type--string files--is used to allow the specification of variable-length byte strings for output to various devices. Since each byte may be either an ASCII character or a binary phoneme code, this file is used to store both ASCII data for eventual output to standard devices (such as the console CRT and printer) and data for output to the voice response units.

The second file type--link files--is used to build three-structure files that represent the desired syntax structure and, in a separate file, the command strings associated with various action triggers. Both these link files are constructed during the input of the system configuration tape. This is in keeping with the overall design goal of allowing application-dependent data to be input from the outside instead of tied directly to the program.

In order to fulfill the requirements of the Word Recognition System several additions were needed to the facilities provided by the aforementioned executive. Perhaps the most far-reaching addition was the inclusion of a 500K word capacity moving-head disk system. Because of the disk, the executive has been modified to operate under the vendor-supplied disk operating system-RDOS. This addition relieved much of the burden of I/O device management and task scheduling from the executive system.

In addition to the disk software, two additional tasks have been added that run in parallel with the system and share certain segments of data. The first of these is the display routine, which is responsible for updating the CRT to represent the current status of each system user. The screen is divided into thirds, with each segment representing one of the three nets. The exact format of the display for each channel depends on the status of the channel at that time, but during normal operation the message entered by the remote user will be reflected on the CRT as it is spoken. The second extra task is operator control. This allows the operator to modify the status of any user and, if necessary, make corrections in the recognized data.

Although these two tasks are independent in time relative to the operation of the executive system, it is obvious that a good deal of information sharing between these processes is required. Once again, this function is provided by the user-state arrays. For example, during recognition the message status is maintained in the appropriate user-state array. Thus, the display routine needs only to access this information to display the proper data, and the operator control task also has access to the same data in order to modify the message.

C. Results of WRS

The requirement for translation accuracy of WRS is 95% over a field radio link with an S/N ratio of 10 dB. This was not acheived. The present design proved unsatisfactory over typical Army radio links. Over an optimal radio link using the international phonetic alphabet, accuracies up to 97% have been observed. Design of the recognition algorithm attempts to minimize WRS sensitivity to amplitude and time variations in the utterance of any word, but tends to be sensitive to normal variations in the speaker's voice, microphone position and background interference.

The critical problem areas appear to be the degree of user training requirements and its impact on overall system reliability. In complex and highly critical waspons and control systems environment for example, an average 95% reliability figure just will not do the job. Neither will complex system user training procedures suffice in terms of multi-user requirement on a potential quick-response requirement. The present Army Word Recognition System requires the user to repeat each vocabulary word approximately five times to develop reference patterns, and complete reference patterns must be stored for each user. With resultant message translation in the 90-95% range in a laboratory environment, the need for additional development effort is readily apparent.

III. SUMMARY AND FUTURE DIRECTION

The application of the speech recognition technology to the Army command/control area presents unique problems as demonstrated by the WRS system. The conflicting requirements for minimum training of the WRS and either a small vocabulary with a broad variation in voice types or a large vocabulary with a limited set of voices present significant complexities to the recognition algorithms. The quality of the Army tactical communications has had a severe impact on the WRS and will force a significant rethinking of the basic approach. The voice communications channel quality is a fact of life that must be overcome if remote voice to computer input is to become a practical reality.

The use of the WRS capability in the console command/control arena eliminates the communications channel problem, but to be useful discrete WRS must provide a much more flexible interaction with a larger vocabulary. The problems of background noise and voice variations remain for this applications area too.

The application of speech recognition technology to tactical military problems introduces a new variable which is unique: stress. What changes take place in the voice input due to stress and how do these effect the recognition process? We could be put in the unacceptable position of having the voice input capability fail us when we need it most.

The capability of speech input and output would be a very valuable one in terms of the man/machine interaction for Army command and control applications. The amount of training the soldier would require in order to interact with and utilize the automated system would be decreased. The man/machine interface would be more "natural" and smoother when it includes keyboard and voice inputs and display and voice outputs. As we have seen from WRS, this is a difficult problem which requires more development effort. The Army is interested in the support of such effort.

ACKNOWLEDGMENTS

The material used for this presentation was taken primarily from the Draft Final Report of the Army Advanced Development Model Word Recognition System (WRS). The WRS program was funded by PM, Army Tactical Data Systems (ARTADS) under the technical direction of Michael Simpson, John Preusse and Merle Hooten. Project personnel at SCOPE Electronics Inc., Reston, Virginia included Joseph C. Brown and Shiraz G. Bhanji.

BIOGRAPHICAL SKETCH

David R. Hadden, Jr.

Education: Reed College, Portland, Oregon, BA (Physics), 1959
University of Pennsylvania, MSE (Computer Science), 1971

1960 to Present: Communications/ADP Laboratory, US Army Electronics Command. Worked in various areas of computer research and development; data storage techniques and technology and computer system design and interface standards. Most recent work has been in the areas of computer terminals and micro-processor applications.