# REAL-TIME INTERACTIVE SPEECH TECHNOLOGY AT THRESHOLD TECHNOLOGY INC.



MARVIN B. HERSCHER

THRESHOLD TECHNOLOGY INC.
DELRAN, NEW JERSEY

## INTRODUCTION

Long recognized as an ultimate step towards simplifying communications between a human and a machine, real-time voice data entry and command and control is now a reality. Over 200 Threshold Technology voice terminals currently are in operation in a variety of industrial, government and military applications in eight countries around the world. To date, over 300 million words and/or phrases have been spoken into these terminals.

Electronic systems are now available which allow a human to verbally input information and/or commands directly into a computer with no keying or handwritten steps involved. Instead of requiring the human to learn the "language" of the machine or the manipulation of special dials or keys, voice input has greatly simplified man/machine communications. With voice input, the operator can provide verbal instructions to the machine in a familiar language which is recognized and translated by the voice terminal to a machine language useful for further processing and/or machine control. Many of the applications using these voice terminals involve some form of interactive feedback from the host computer to the human operator. Consequently, system design involves more than simple speech recognition and must consider a variety of man/machine interactive relationships. Performance achieved in the laboratory by highly motivated personnel usually cannot be achieved in "real-world" environments by less motivated individuals unless a variety of human factors are considered.

Threshold Technology has had voice input systems operating in these "real-world" environments since late 1972 and has gained a great deal of knowledge regarding user acceptance and human factor requirements. Based upon this experience, improvements in speech recognition techniques (involving both hardware and software) have been evolutionary, and the interactive relationships between the operator and the machine have continually been improved.

In most applications, the systems are highly interactive in that information is displayed to the operator denoting his next input requirement or showing the last recognition decision. Visual and/or

audio verification and the ability to edit the verbal input can produce virtually error-free data input. In this manner, the voice entry system has been designed around the requirements of the human, thereby greatly simplifying the task of man/machine communications.

Additional aids often can be provided to the operator to assist him in his use of the speech recognition system. These include a wireless input to permit operator mobility and a remote input console to provide a simplified means of accessing speaker reference data or changing vocabulary words.

This paper will review the basic real-time isolated-word recognition techniques developed by Threshold Technology, together with some of the commercial products employing the techniques. Some of the industrial applications will be reviewed which serve both as a chronological history of the application of this equipment, as well as an illustration of the diverse usage. Next, some of the prior and current Government supported R&D efforts will be discussed, along with the qualifications of technical personnel and our general and special facilities.

## THRESHOLD RECOGNITION SYSTEM CONSIDERATIONS

### General

The Threshold Technology recognition equipment is a speaker adaptive, real-time, isolated-word recognition system. Isolated word recognition requires that there be a short pause before and after utterances that are to be recognized. The minimum duration of the pause is on the order of 100 milliseconds in order to minimize the confusion which might arise due to stop gaps appearing within the utterance. Although it is more natural not to require pauses between words, it should be pointed out that most practical applications can be satisfied using isolated-word systems, and that this restriction has not presented user training problems. Industrial workers have readily adapted to speaking words in isolation and have achieved speaking rates in excess of 70 words/minute for sustained periods of time with peak speaking rates in the area of 120 words/minute.

Most practical applications require a vocabulary consisting of about 20 to 30 words, but the Threshold speech recognition terminals can easily be made to handle 200 or more words or short phrases simply by adding to the modularly expandable memory of the speech recognition processor. The entire system is programmable such that individual words can be changed or the whole vocabulary and syntax structure can be changed, depending on the application.

A block diagram of the speech recognition system is shown in Figure 1.

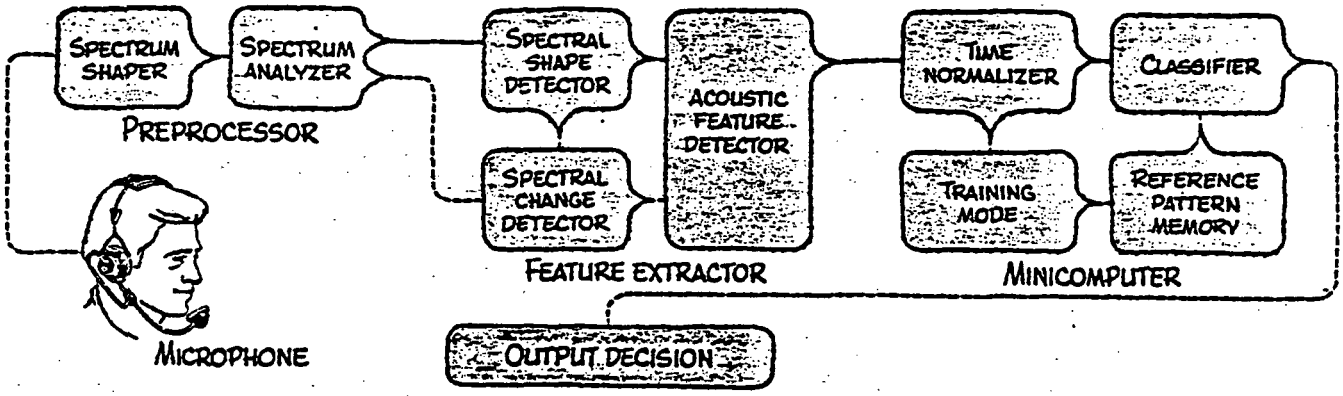


Figure 1. Block Diagram of Speech Recognition System

## Preprocessor

One purpose of the preprocessor is to shape the output from the noise-cancelling microphone to remove irregularities and produce a normalized speech spectrum. This equalized signal is then passed through a real-time spectrum analyzer consisting of a contiguous back of active bandpass filters. Originally, 19 filters were used in the VIP-100, ranging in center frequency from 260 Hz to 7626 Hz. Currently, 16 filters ranging from 260 Hz to 4484 Hz are used in the new Threshold preprocessors. The outputs of these filters are full-wave rectified and logarithmically compressed. This latter operation provides a 50dB dynamic range and produces ratio measurements when subsequent features are derived from summation and differencing operations, thereby minimizing the input amplitude dependence.

## Feature Extractor

The function of the spectral shape detector is to develop spectral derivative (dE/df) features indicating the overall spectrum shape. The spectral shape and its changes with time are continuously measured over the frequency range of interest. Combinations and sequences of these measurements are processed in hardware to produce a set of 32 significant acoustic features, one of which is the initial estimate of word boundary. A more refined word boundary is derived in the computer using the variable backup technique.

## Recognition Accuracy

Error rates in a practical speech recognition system must be sufficiently low to eliminate any loss of operator confidence or efficiency. Humans have a tendency to become oblivious to very low error rates in multitask operations. Corrections by voice must be sufficiently infrequent as to be no hindrance to the accomplishment of the intended task. If the error rate is high enough to interfere noticeably with the task, the operator will lose confidence and will not wish to use the voice input system. In a sense, the operator makes a binary decision, i.e., the voice input system is either "good" or "bad". Interviews with users of operational voice input systems have shown that rarely is a voice input system accepted unless the error rate is very low. The acceptable error rate is critically dependent upon the particular application and the data entry rate. In most practical applications, the "raw" recognition error rate usually must be less than 1-2%. High voice data entry rates about 50 words or phrases per minute require lower error rates than those applications where the data rate is slow enough that the user has time to make corrections.

Variations in speech patterns are found even when the same person repeats a word, particularly over a period of time. This complexity is greatly magnified when different speakers say the same word. Such differences have made the design of accurate "universal" recognition systems a formidable task. Consequently, almost all systems (including Threshold Technology) now in practical use employ the speaker adaptation design. Once a speaker "trains" the machine - by repeating each word in the vocabulary approximately ten times - the parameters of that voice are stored permanently in the system's memory. At the start of operation when an operator comes on duty, he simply inputs his code number into our remote console and the vocabulary information he has previously recorded is automatically transferred to the active memory system.

## Background and Breath Noise

Background noise can be a real problem in many applications where these systems may have to operate at a noisy work site. A contact microphone does not solve the problem because it would also cancel some of the attributes of unvoiced frictional sounds, making recognition more difficult. The contact mike can also produce erroneous signals that are the result of body movement. Head movement away from a highly directional microphone causes wide frequency variations which also would make speech recognition difficult. The most practical compromise has been the use of a noise-cancelling microphone on a lightweight headband. It maximizes the signal-to-noise ratio, moves with the speaker, and frees the operator's eyes and hands for other related tasks.

Breath noise becomes a serious problem with a closely-mounted microphone. Exhaling can produce signal levels in a microphone comparable to speech levels. To separate speech information from breath noise, each of which does have unique spectral characteristics, the Threshold system utilizes pattern recognition processing which discriminates between speech sounds and the frictional breath noise.

Word Boundary Detection

For isolated word recognition, accurate word boundary detection is very important. Word boundary detection initially is derived from a combination of the overall amplitude of the speech, together with information obtained within predetermined spectral bands. This boundary signal input is dampened enough so that it will not react to brief intervocalic pauses caused by stop consonants and affricatives. A variable back-up boundary duration is used, together with the breath noise detection, to isolate the speech information. The boundary detection must also be accurate even when background noise is high.

Operator Babble

Since these isolated-word recognition systems recognize a limited vocabulary, it is important to minimize false recognitions of utterances or sounds not included in this vocabulary. Operator-originated babble is inevitable as it can be caused by coughs, sneezes, throat clearings or side conversations occurring when the operator forgets to turn off the microphone. These types of sounds ideally should be rejected by the recognition system. In the Threshold equipment, a rejection criteria is derived and either an audio and/or visual feedback message is given the operator when an input utterance is not accepted by the system. Another safeguard to prevent inadvertent message entry to the speech recognition system is to employ syntax and to format the data entry sequence as much as possible so that after a block of data has been entered, a verification word is required before entry is considered to be valid by the speech recognition system.

Feedback, Editing, and Interaction

Immediate feedback must be given the user of a voice input system, either visually, aurally, or both. The feedback must be unambiguous and can greatly assist the user in accomplishing his voice input functions.

In an isolated speech recognition system, it is important to pace the user so that the minimum spacing is maintained and words are not run together. This can be achieved by an audible "ready" tone or visual indicator. An experienced user of an isolated word voice input system will quickly learn the fastest rate at which words can be spoken, after which the "ready" indicators are unnecessary. However, in the initial stages of using a voice input system the ready indicator is a valuable training aid to the operator.

A "reject" indicator similar to the "ready" indication can also be useful as discussed earlier. The reject indication may also serve the purpose of subconsciously training the operator to speak the vocabulary words used in a manner that can most easily be recognized by the system.

Besides the elementary indications of "reject" and "ready", all spoken commands should be fed back to the operator for verification. This verification can take the form of a positive indication of correctness through a control word such as "OK" spoken after each command or each data field, or can simply be indicated by proceeding to the next command. Control words such as "erase" to delete the last command and "cancel" to delete an entire data block should also be provided.

It is in the area of conversational, interactive feedback that the greatest potential exists to assist the user of a voice input system. Feedback to the operator cannot only be used for verification, but also for prompting the user through an entry sequence, checking syntax, format and expected values and making special inquiries when the application requires such.

### Stability of Reference Data

As mentioned previously, an adaptive, limited vocabulary system achieves recognition processing by comparing an unknown utterance with a set of stored samples of the vocabulary words obtained from the user of the system. This reference data must be stable over long periods of time for practical applications. Once the reference data has been obtained the operator should be able to use the voice input system with little or no "retraining". The ability to begin operations each day with no "warm up" or retrain will greatly enhance the operator's confidence. Similarly, he should not have to frequently interrupt his normal operations to retrain individual words during the course of operations.

## THRESHOLD RECOGNITION SYSTEM

### Description

The basic speech recognition system was described at the 1972 IEEE Conference on Speech Communication and Processing[1]. This initial limited vocabulary system was designated the VIP-100 and consisted of a hardware speech preprocessor and feature extractor, together with a classifier function performed by a minicomputer. The minicomputer also time normalizes word durations, performs adaption to new talkers and/or words during the training mode, and provides storage of the reference patterns for each word. The minicomputer originally used was a Digital Equipment PDP-11, later changed to a Data General Nova 1200, Nova 2 and Nova 3 as the Nova family evolved. A current version of this system (designated the Threshold 500) utilizes a Digital Equipment microcomputer - the LSI-11 - in place of a minicomputer.

The features used in the recognition system are a selected subset (including complex combinations) of acoustic features functionally similar to those described in Reference 2. Each feature is extracted by a combination of analog operations and binary logic. The output of the feature extractor consists of 32 binary signals, $F_1$, $F_2$ ... $F_{32}$. These features are of two types, 16 broad-class features and 16 phonetic event features. The broad class features include such categories as vowel/vowel-like, formant characteristics, short pause (less than 100 ms) and unvoiced noise-like consonant. The 16 phonetic event features represent measurements corresponding to phoneme-like occurrences.

Classifier

This portion of the Threshold recognition system includes a time normalizer, training mode, reference pattern memory and a decision algorithm. A general-purpose minicomputer or microcomputer is used for these functions.

The 32 encoded features and their times of occurrence are stored in a short-term memory. When the end of the utterance is detected, the length of the word is computed and divided into 16 equal time segments and the features are reconstructed into a normalized time base. The pattern-matching logic subsequently compares these feature occurrence patterns to the stored reference patterns for the various preset vocabulary words and determines the "best fit" for a word decision.

A total of 512 bits of information (16 time segments, each containing 32 features) are required to store the feature map for each utterance or reference pattern. For a thirty-two word vocabulary, the information stored requires 16,384 bits. Since minicomputers operate at 0.2-0.5 mips (million instructions per second), the response time is immediate for small vocabularies. For larger vocabularies, a separate hardware high speech pattern-matching comparator is employed to minimize response time.

Training Mode

The voice system, being adaptive, requires "training" for individual talkers and/or words. The system can be automatically "tuned" to the voice characteristics of any single user in a short time period simply by speaking each desired word approximately 10 times to provide a reference set of features. The system stores in memory an individual reference set of word features for each word in the vocabulary and for each talker in the system. Once having trained the system, new words spoken into the device during normal operation are compared with the stored references and a "closest fit" is selected as the recognized word. It is also possible to obtain a "no-decision" or reject when none of the characteristics of the words in the reference memory are close to the spoken word.

In training the machine, the system automatically extracts a time-normalized feature matrix for each repetition of a given word. A consistent matrix of feature occurrences (between repetitions) is required before the features are stored in the reference pattern memory. A template threshold factor is chosen such that a feature occurrence (in a given time segment) is considered valid only when it occurs a minimum number of times relative to the number of training samples. Usually, this threshold factor is set to be between 30-50% of a feature's occurrences within the samples. An example of a reference feature matrix and a test word matrix for the word "seven" is shown in Figure 2.

Recognition Mode

Once the parameters of recognition are set, a spoken word is digitally compared to each stored reference matrix. Similarities and dissimilarities in each compared matrix are appropriately weighted and the net result provides a weighted correlation product. Correlation products also are generated after shifting the input word matrix ± one time segment. The stored reference word producing the highest overall correlation is selected as being correct, providing it exceeds a minimum correlation threshold value. The references used by the system are normally not effected by operator abnormalities such as head colds, sore throats and hoarseness.

## THRESHOLD 500 VOICE DATA ENTRY TERMINAL

The microcomputer-based voice data entry terminal - the Threshold 500 - was first introduced commercially by Threshold in late 1975. The Threshold 500 voice data entry terminal normally operates in conjunction with a host computer system. In this configuration, a system is capable of handling multiple talkers and multiple input terminals. Each terminal can accept voice input, produce a recognition decision, drive a display and interface to other equipment. In essence, each voice data entry terminal may be considered a computer peripheral capable of performing independently as a data entry device. The Threshold 500 system can be software configured as a standard keyboard replacement or a sophisticated, interactive, intelligent terminal with local processing.

Figure 3 is a block diagram of a typical multiterminal Threshold 500 system. In this configuration, the central computer (which could be a minicomputer) acts as a system controller which can accept input data, transmit and receive speaker reference data, and control the display messages associated with each terminal. Asynchronous serial communication with each Threshold 500 is utilized for these purposes. A disk file often is provided which can be used to store the data base as well as speaker reference data. Reports and statistics related to a particular application can be generated and printed out.

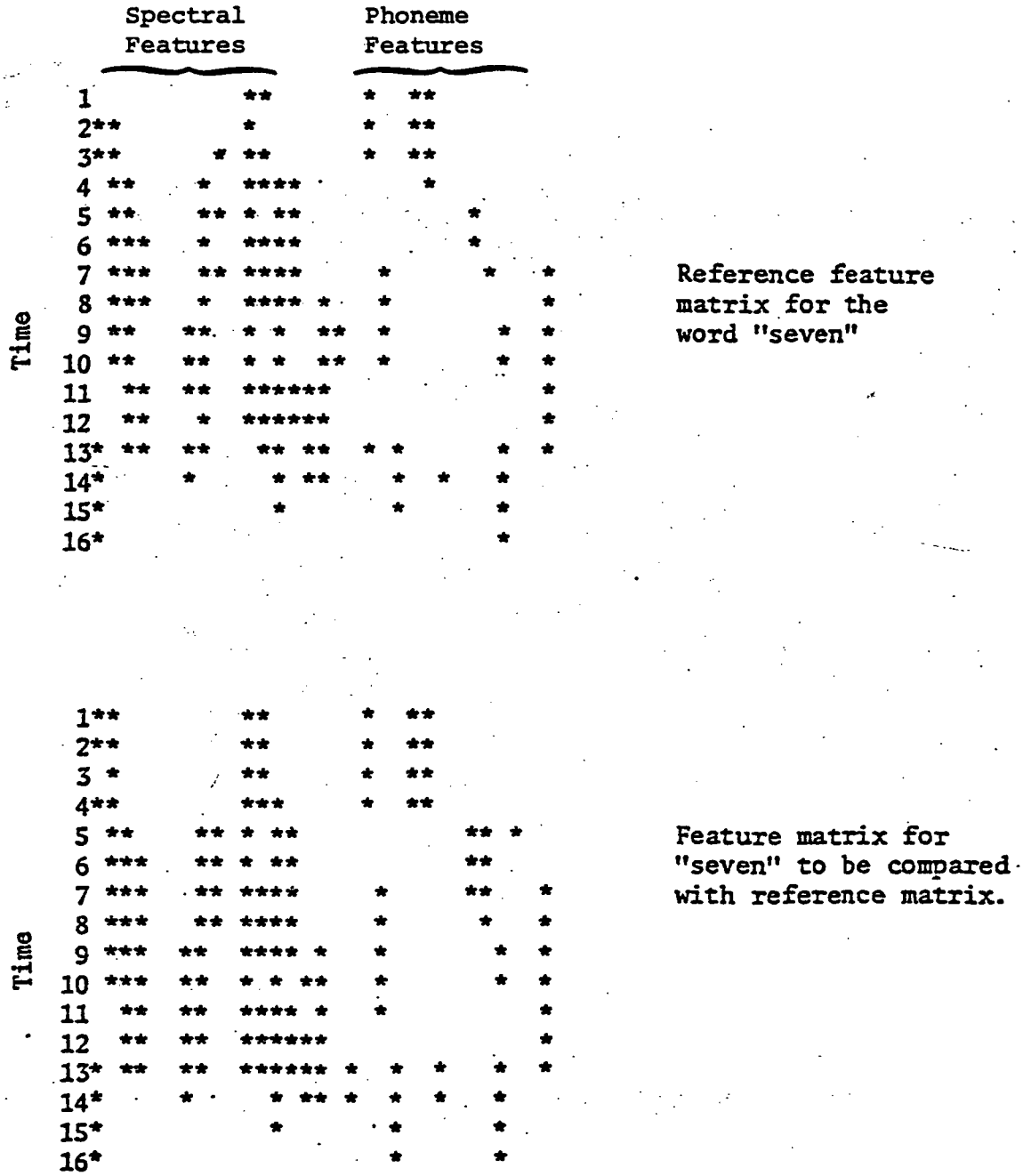


Figure 2. Sample Reference and Test Matrices for the Word "Seven"

A standard Threshold 500 terminal includes a recognition subsystem, a display and a remote operator console. The output of each voice data entry terminal is in ASCII code, each word or phrase recognized by the terminal producing a unique character. This output is configured for EIA RS232C, CCITT-V24, or 20 mA current loop teleprinter compatible. Full duplex communications are provided and data transmission can be made character-by-character or by a verified data field. Consequently, the Threshold 500 can be linked easily with a central processor to provide voice input in place of keyboard data or other entry devices. The standard terminal employs a volatile semiconductor memory which can be down loaded with operator reference data from a central file via a communications line or trained locally by providing spoken samples of the desired words or phrases. A core memory can be substituted in the terminal to provide a non-volatile memory for speaker reference data. The control and speech processing software is stored permanently in the terminal in a semiconductor read-only-memory.

Vocabulary words can be trained or retrained locally and different operators can use the terminal by selecting the appropriate word numbers and/or operator numbers on the local operator control console. A local interactive visual display permits prompt messages and recognition results to be displayed to the operator. A special communication protocol has been developed to transmit operator reference data and output decisions to and from the host computer.

## CRT COMPATIBLE VOICE DATA ENTRY TERMINAL

In some applications, the requirement for storing operator reference data in the host computer and transmitting these data to and from the terminal is not desirable. Also, handling of the special protocol to allow the transmission of both ASCII characters and the binary speech reference data can unnecessarily complicate the software required in the host computer. For these applications and to minimize the programming required to use the Threshold 500, a new CRT/teleprinter compatible voice data entry terminal, designated the Threshold 600, has been developed. This new terminal is plug compatible and transparent with all asynchronous CRT's, terminals and Teletype®-like devices. Consequently, this new voice data entry terminal can be interchanged directly with an existing terminal attached to a minicomputer, a time sharing system or even a large computer providing asynchronous terminal support exists.

The main reason this compatibility can be achieved is the incorporation of local storage at the terminal for operator reference data. Consequently, reference data need not be transmitted and stored at the host computer. Several additional attributes can be achieved through the employment of local storage. With an auxiliary keyboard, training message prompts also can be locally defined and stored by the user.

In addition, output messages and/or control functions also can be locally defined and stored. These output messages can be single ASCII characters or strings of characters as defined by the user.

Consequently, the asynchronous CRT compatible Threshold 600 terminal permits the user to program which words and/or phrases he wishes to have the terminal recognize, as well as what characters he wishes to display and transmit to the host computer. This user programmability means that this same terminal can be used for a variety of applications since individual programs can be recorded for each application and read into the terminal when required. As an example, the user may wish to define one of the vocabulary words as the common "rubout" function used in many teleprinter applications. The training prompt can be defined, by auxiliary keyboard input, in the programming mode to be "RUBOUT" or "ERASE" or "DELETE", etc. The output can also be defined as the ASCII "DEL" character. Thus, when the operator first trains the terminal to recognize his utterances, he will be prompted by the local terminal display (without host computer intervention) to say "RUBOUT" or "ERASE" or "DELETE". When the operator subsequently speaks that word in the recognition mode, the "DEL" character will be sent to the host computer.

The keyboard used for user programming can be supplied with the terminal or can be any other asynchronous keyboard terminal. Also, if need be, the prompts and output characters can be down-loaded into the terminal memory from the host computer.

An optional line buffered mode is available which allows local control of functions such as TRANSMIT, DELETE, etc. Consequently, in this mode, intelligence is added and local editing functions can be achieved, minimizing the burden on the host computer. Other optional functions also are available which can further increase the effectiveness of voice data entry beyond that achieved in the standard keyboard terminal.

## APPLICATIONS

The potential applications of voice input for data entry and command and control are enormous. Commercially at the present time, these applications are limited mainly by the economics involved in cost justifying the replacement of existing alternative data entry devices and/or techniques. As the cost of voice terminals decrease, more justifiable applications will arise, particularly when the true costs of data capture are considered. In some cases, voice input offers advantages which far outweigh the direct labor savings and permit certain operations to be achieved which could not easily be accomplished using alternative data input techniques. This is particularly true in

certain potential military applications. Table I is a summary of some of the current applications of Threshold voice input terminals.

A brief chronological history of the installation of Threshold recognition systems for selected industrial applications illustrates the diverse useage and some of the advantages obtained by the use of voice input. Descriptions of additional industrial applications of Threshold systems are presented in Reference 3. Some of the Government applications and R&D efforts are described separately in a following section.

TABLE I

CURRENT APPLICATIONS OF THRESHOLD VOICE INPUT TERMINALS

MANUFACTURING AND DISTRIBUTION

- Factory Source Data Collection
- Quality Control and Inspection
- Parts Programming for Numerically Controlled Machine Tools
- Receiving, Shipping and Inventory Control
- Material Handling and Sortation Systems
- Production and Process Control
- Industrial Robots and Machine Control
- Computer-aided Design

VOICE DATA ENTRY

- Keyboard Replacements
- Financial Reporting
- Intelligent Interactive Terminals

GOVERNMENT

- Air Traffic Control
- Cockpit Control
- Shipboard Fire Control
- Aids for the Handicapped
- Cartographic and Hydrographic Data Entry
- Computer-aided Instruction

## Material Handling

The first speech recognition terminal was installed by Threshold Technology in a commercial environment late in 1972 for an airline baggage handling application. This type of material handling application permits the simultaneous handling of parcels or bags and the entry by voice of a destination code to operate a mechanized conveyor system. One man, using a voice encoder can control the conveyor delivery system thereby eliminating a second operator who formerly was used to key in this information. These systems also can provide piece counts for individual operators and other statistical data in printed report form. Since the initial installation, many additional systems have been installed at various airline, retail distribution center, and parcel delivery service locations. One such system currently operating has the capability of accepting 42 simultaneous inputs using the Threshold 500.

## Inspection and Quality Control

The second speech input system was installed in January 1973 at Owens-Illinois for voice input of product inspection data directly into a computer, providing an automatic hard-copy printout of the results. This system has been operating 3 shifts a day, 7 days a week since installation and is quite typical of many of the inspection and quality control systems subsequently installed. Using the voice data entry system an inspector can enter his (or her) data simultaneously with the inspection and thereby increase overall productivity. The voice data entry system is programmed such that the inspector can simply follow a checklist appearing, item by item, on an electronic display, and enter measurements via a "hands-free" operation while visually verifying that the information was correctly accepted by the system. Errors can be corrected using a control word such as "Erase" and corrected data re-entered. Consequently, 100% correct data can be entered into the data collection system with no time delay or errors of the types associated with inspection techniques using manual recording or keying. In this application, as well as many other installations, measurement tolerance data can be stored in the computer and the operator alerted when input measurements are out of tolerance. Various types of reports can be generated since all of the data from the operators is recorded by the system and statistical summaries of the results can be printed out or displayed on a CRT.

More advanced quality control systems have been installed in various can manufacturing plants throughout the country to assist the manufacturers in maintaining the quality of their products. These systems use multiple Threshold 500 terminals operating in a mode similar to that illustrated in Figure 3.

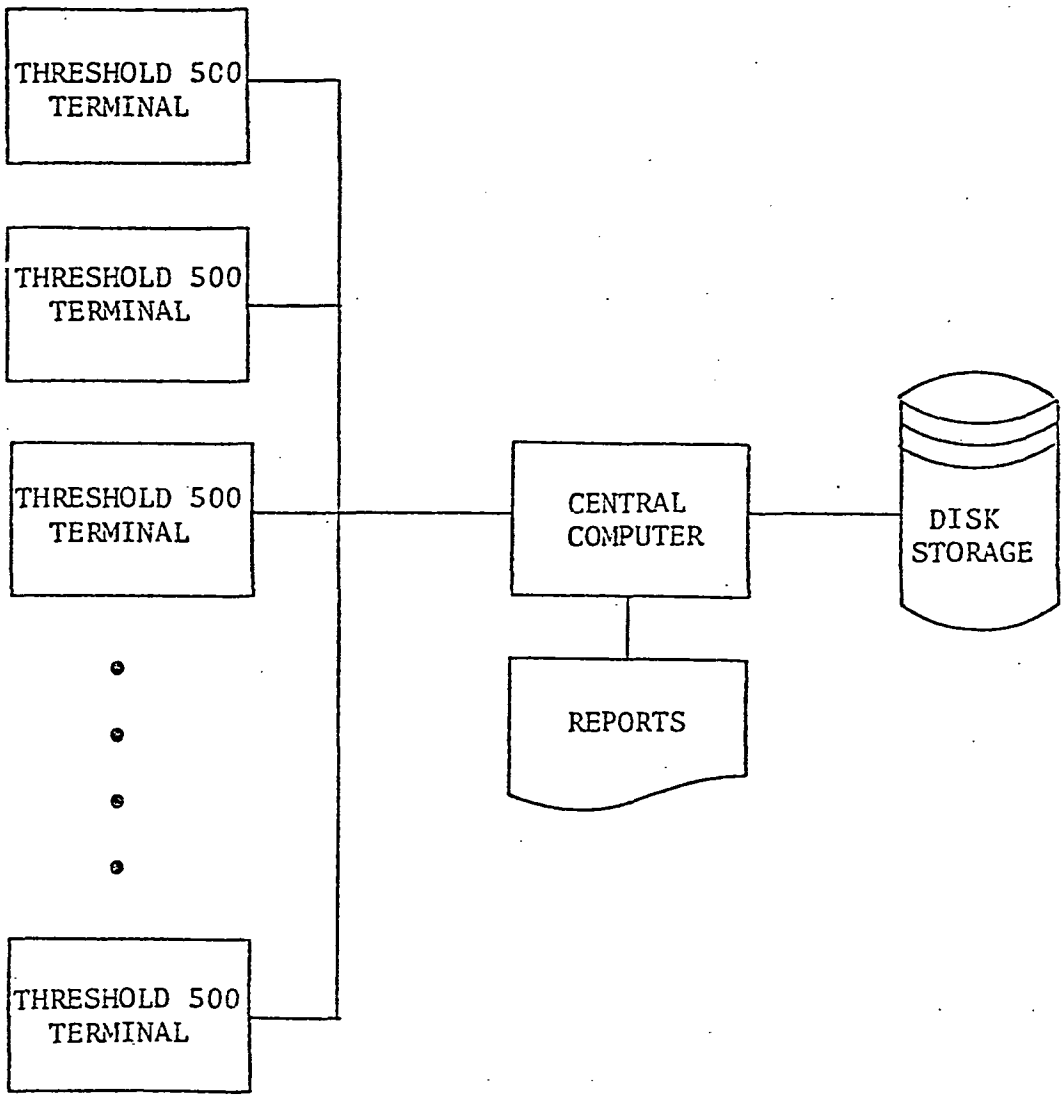


Figure 3. Threshold 500 Terminal System - Block Diagram

## Source Data Entry

A typical installation of a voice input system exclusively used for source data entry was made in 1974 at Tecumseh Products Company. In this receiving application, compressors returned to Tecumseh for service analysis required the preparation of a form for each compressor. This operation necessitated writing down such items as order number, item number, complete serial plate data, customer tag numbers, etc. Handling the compressors and writing down the information on receiving forms was both time-consuming and error-prone. These forms then had to be keypunched by the data processing department for computer entry which led to further time delays and errors.

The use of a voice input system for direct data entry to a computer overcame all of these problems. The operator now speaks the data as he handles each compressor and is guided through the entry sequence by a display. Erroneous serial plate codes are spotted as the operator enters the data. Thus, immediate correction is possible at the time of recipt of the compressor rather than after the compressor has been sent to a repair area. This type system has increased operator productivity and accuracy and considerably reduced response time to customer inquiries.

A variety of other types of data entry system applications also have been installed. These systems range from the entry of financial information for consolidating balance sheets to the entering of stock and bond transactions via voice as well as recording serial numbers for product information collection.

## Machine Tool Programming

A fourth major application area for speech input is programming numerically controlled (NC) machines in the metal-working industry[4]. A traditional obstacle to the use of computer-based numerical control systems has been the human interface problems associated with programming and software. The use of voice programming has made it possible for machine shop personnel, relatively unfamiliar with programming languages, to prepare fully verified punched paper tape programs for a variety of automatic machine tools. The programmer simply speaks into a microphone each programming command in sequence, using normal English words, and the system automatically "decodes" the information into a machine-compatible format. As part of this system, a display not only flashes each command spoken to provide instant, positive verification or correction, but also displays the next entry required, thereby interactively sequencing the operator through all of the steps necessary to produce a program tape for any particular NC operation.

This type of system has been designated a VNC (Voice Numerical Control) system. This family of equiment represents the first practical example of computer programming via voice input, and appears to provide the ultimate in simplified communications between man and production machines. The first VNC system was installed early in 1975 and additional more advanced versions subsequently have been installed.

## GOVERNMENT APPLICATIONS AND R&D

### General

Threshold Technology personnel have conducted a variety of Government sponsored R&D programs involving real-time speech recognition, speaker authentication and identification, keyword recognition, and language identification. In addition to study programs, many of these R&D projects involved the delivery of real-time recognition hardware for further evaluation. In some cases, these equipments are being used by Government personnel in actual operational environments. Additionally, standard and modified speech recognition equipment manufactured by Threshold have been delivered to various Government and commercial activities either for incorporation into larger systems for military applications or for in-house experimentation by Government facilities. Several of the applications of some of these delivered equipments will be briefly described.

### Voice Control Demonstration System for Cockpit Functions

In early 1974, an experimental speech recognition system was delivered to the Air Force which was to be used to demonstrate voice control of aircraft cockpit functions. The Voice Control System developed during this contract was a self-contained, real-time isolated word recognition system designed to recognize a limited vocabulary of 144 words. The system could be used by either of two operators at any given time. The adaptive system could be retrained quickly for new vocabulary words or other operators. Operational flexibility was achieved through the use of a variable command format structure, under program control. Figure 4 shows the syntax designed into the Voice Control System Command structure. The system was designed to operate with either a standard noise-cancelling microphone or the integral M-100 microphone of the MBU-5 oxygen mask. A digit recognition accuracy of 99.79 per cent was obtained for ten speakers using a standard noise-cancelling microphone in a laboratory environment. Recognition accuracy for non-digits was 99.32 per cent under the same conditions. An overall recognition accuracy of 97.15 per cent was achieved with the M-100 microphone in the laboratory environment with the subjects breathing compressed air or oxygen through the MBU-5 oxygen mask. The results obtained during this program were promising and indicate that additional

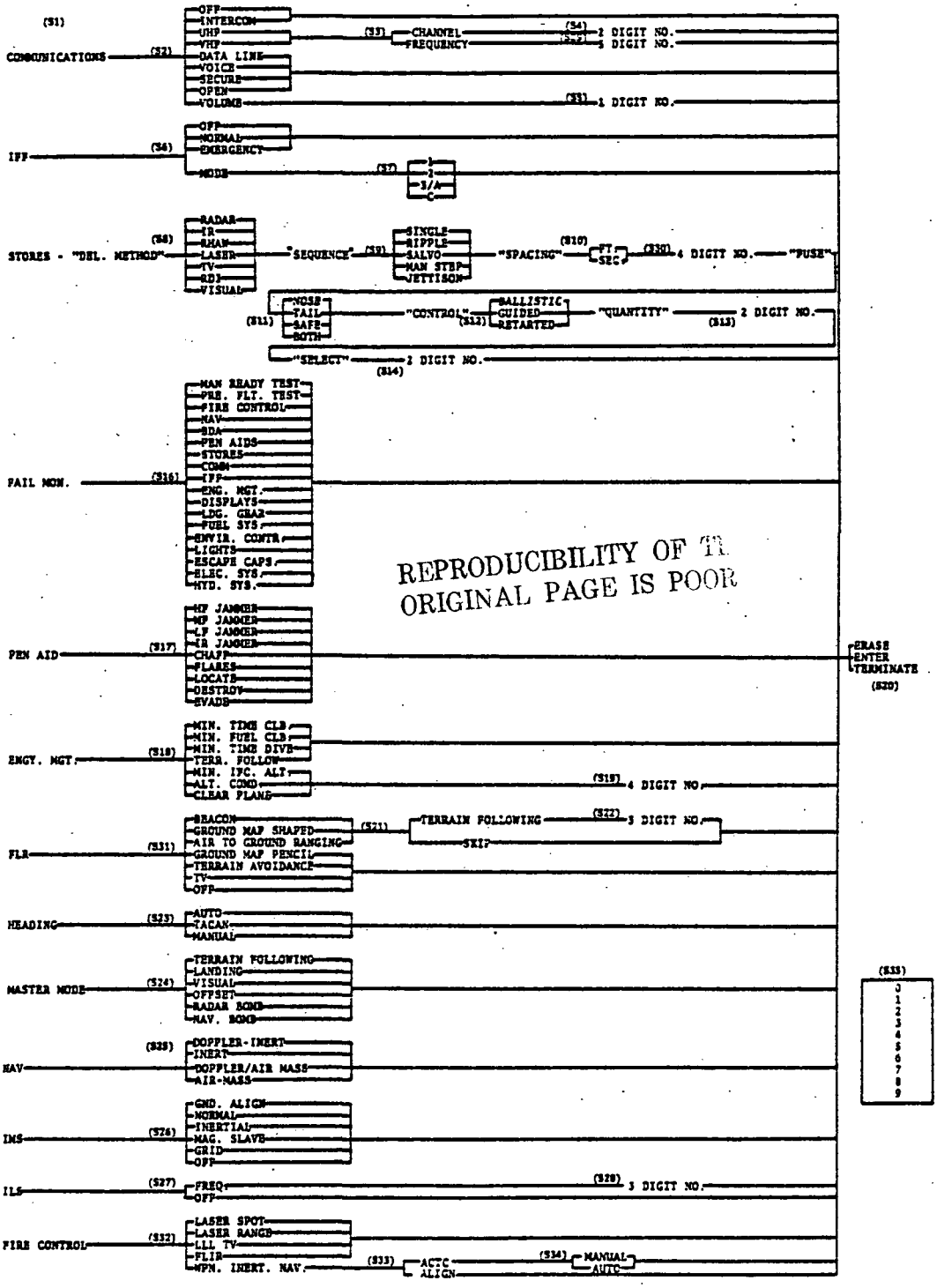




Figure 4. Voice Control System Command Structure

studies should be undertaken to determine the effects that altitude, "g" forces, and operator stress would have upon recognition accuracy in order to ascertain the operational feasibility of the application.

Cartographic Data Entry

Cartographic data entry can be simplified by the use of speech recognition equipment. In preparing maps, the cartographer normally has to look up from the map to manually key the data into the computer. This interrupts the procedure, reducing efficiency and increasing the chance for error in the eye movement from keyboard to map and back. With voice input, the cartographer can speak the information directly into the computer without stopping and with his hands and eyes remaining on the source of the information.

In 1976, Threshold delivered a VIP-100 recognition system to RADC for cartographic use. This equipment was originally intended to be used with a bathymetric digitizing system located at RADC to input bathymetric depth readings from smooth sheets. In early 1977, this system was moved to the Defense Mapping Agency Hydrographic Center (DMAHC) and interfaced with a bathymetric digitizing table. The resulting combination provides an improved means for entering bathymetric readings from smooth sheets directly to punched cards. The system allows the cartographer to simultaneously obtain X-Y coordinate locations and provide voice data entry of depth reading for each coordinate location. With the operator's hands free to concentrate on the X-Y position sensing device (cursor), the operator can speak the depth number. These readings can be verified using a small LED display mounted on the cursor, and if they are correctly recognized, he or she can enter them directly onto punched cards without losing sight of the smooth sheet. The system at DMAHC has a vocabulary of the ten digits plus four control words. It can store reference data for five speakers. A special provision was made to allow correction of previously entered depth data. Any of the last five depth entries can be corrected at any time. Cards containing X, Y and Z data are punched automatically.

Also developed during this program was an advanced development model of a highly reliable isolated-word, speaker dependent, limited-vocabulary word recognition system based on the VIP-100. This system recognizes up to 200 words arranged in a structured manner. The structure consists of any combination of nodes. Each node can include up to 30 vocabulary words. A multiplicity of node plans can be stored on the system, together with speaker reference data for up to 20 speakers. An advanced development system which can handle up to 600 words currently is being developed for RADC under a different program which will assist cartographers in inputting data for application to a Digital Radar Landmass Simulator (DRLMS) and for production of Flight Information Publications (FLIPS).

## Air Traffic Control

Threshold speech recognition systems also are being investigated for various applications related to air traffic control. In June 1974, a VIP-100 system was delivered to the Naval Training Equipment Center (NTEC) for incorporation in a system to train Ground Controlled Approach (GCA) controllers. The overall GCA training system was developed by NTEC and Logicon, Inc.[5] Additional Threshold preprocessors currently are being incorporated by Logicon, Inc. into Automated Adaptive Flight Training System (AFTS) applications. The AFTS works in conjunction with existing flight simulators to automate the training syllabus associated with Instrument Flight Manuevers (IFM), GCA, Air-to-Air Intercepts (AAI) and Ground Attack Radar (GAR) operations. The AFTS has been developed and integrated by Logicon into F-4E and TA-4J flight simulators.

Another VIP-100 system was delivered to the FAA-NAFEC in May 1975 for experimentation in actual air traffic control applications. Controllers, in addition to their monitoring, managing and decision-making tasks, often have to type into a computer the instructions transmitted to a pilot by voice. Speech recognition equipment could accomplish both the pilot instruction and computer up-dating with the same voice transmission, allowing traffic controllers to keep their attention on the monitoring equipment. Dr. Connolly, of NAFEC, in a separate paper, will describe some of the possible applications and experimental results, to date.

## Voice Input Code Identifier

A Voice Input Code Identifier (VICI) advanced development model was delivered to the Air Force (RADC) in early 1975. VICI is an isolated word speaker-independent recognition system capable of recognizing the English digits and four control words, CANCEL, ERASE, VERIFY and TERMINATE. By the use of an alphanumeric output display, a speaker using the system can verify that each digit spoken into the system was correctly recognized. Errors can be corrected through the use of the control words. The VICI system is based upon the VIP-100 isolated word recognition system which normally requires the input of training data by each talker who uses the system. For use in the VICI application, both hardware and software modifications were made to a VIP-100 system to allow recognition of the VICI vocabulary spoken by a large speaker population without adaptation or training by any speaker.

VICI was developed to fulfill a requirement of the Air Force Base and Installation Security System (BISS). BISS requires a completely voice-oriented technique for a person entering an Air Force Base to claim his identity and be verified. Such a technique would

eliminate the need for picture badges, keypunching code numbers, and other fallible mechanical methods of entering an identification number. The speaker would simply utter his code numbers (sequence of four digits and one or two check digits) to VICI and if correctly entered into the system, automatic speaker verification could then be performed by another subsystem.

The original VICI system was developed for use by male talkers only and required wide bandwidth speech data input. In 1976, a subsequent R&D program modified the system such that it could recognize the same 14 words spoken over telephone line bandwidths by either males or females. Also provided was an error detection/correction scheme using 2 check digits to minimize code number entry errors. The system corrects code errors when possible or requests a reentry of the data by the talker.

For the wide bandwidth, male talker only system, performance was as follows. Individual digit recognition accuracy in each of two tests from magnetic tape was 98.7% for a total of 65 speakers. In live tests, a total of 30 speakers each spoke 75 groups of digits, each group consisting of four digits followed by the word VERIFY to simulate operational conditions. Individual digit accuracy in these tests was 97.9% for 30 speakers. Approximately 92.5% of all digit groups were inputted and verified without error. (No check digits were employed in these tests.) The remaining groups were corrected and properly entered. With feedback verification and error correction, all talkers in the live tests were able to enter all digit groups correctly. Most codes, together with the verify command, were entered in four to seven seconds when no errors were detected.

The telephone bandwidth system was tested with a total of over 56,000 words spoken by both male and female talkers. Individual digit accuracy in the tests conducted by the use of tape recordings was 96.85% for 182 talkers. All tape recorded data were passed over actual telephone loops which included two centrals and a connecting trunk as well as lines to and from centrals. Limited testing of the error detection/correction scheme involving 29 talkers indicated that 54% of the incorrect code groups (4 digits) could be corrected automatically.

Aids for Handicapped

Perhaps one of the most humanitarian aspects of Threshold speech recognition systems is as an aid to handicapped individuals. With speech recognition, a severely disabled person can be given control of his environment. Voice-controlled wheelchairs, beds, typewriters, telephones, calculators and servomechanisms are all possible.

Prototype systems have already been developed for this application. One such system has been delivered to the Veterans Administration (VA) and currently is being tested at a hospital.[6] This system provides (1) a voice activated environmental control unit, (2) a typewriter input/output, (3) a four-function calculator with memory, and (4) a telephone dialer. Another system has been built to operate a wheelchair. Currently, a system is being developed for the VA to operate a wheelchair as well as an attached extendable mechanical arm, both via voice control.

## THRESHOLD PERSONNEL AND FACILITIES

The technical personnel at Threshold have had a long history of performing R&D work. These R&D programs include both Government sponsored as well as in-house supported efforts. It is important to note the fact that the only business of Threshold Technology is the development of products utilizing speech recognition and processing. Currently, 12 professional and technical personnel are involved in these speech related activities. Collectively, these personnel have almost 100 man-years of expertise in the field of speech processing and recognition. These engineers have directly contributed to and/or managed over six million dollars of in-house and government sponsored R&D efforts in the speech area over a period of 17 years. A summary of some of the achievements of Threshold personnel in speech recognition is shown in Table II.

Although most efforts at Threshold Technology have been in the development of real-time isolated word and connected word speech recognition systems, extensive work has been performed in speaker authentication and identification, keyword recognition, language identification, and speech bandwidth compression. Additionally, a recent contract with the Air Force (RADC) involved a study to perform an analysis and an experimental evaluation of human factors and other problems associated with inputting data into an information data handling system. The input modes studied included voice and several other manual modes. Measurements were made of efficiency and accuracy, and an assessment was made of the various devices' applicabilities to future man-machine interfaces.

### Facilities

Threshold Technology Inc. occupies 18,000 square feet of a single story of a modern facility. In addition to a variety of standard laboratory test equipment, a 12 channel and a 20 channel optical oscillograph are available for the simultaneous parallel analysis of speech features. Since we manufacture speech recognition systems, a

## TABLE II

## SPEECH RECOGNITION ACHIEVEMENTS

| | |
|---|---|
| 1960 )<br>1961 ) | Invented hybrid logic suitable for real-time pattern recognition (analog-threshold logic). |
| 1962 )<br>1963 )<br>1964 | Demonstrated vowel recognition )<br>)<br>Demonstrated consonant recognition ) Isolated speech. Obtained highest reported accuracy. |
| 1965 | Demonstrated feasibility of recognizing continuous speech. Invented basic speech synthesis technique. Constructed and delivered speech-recognition system for Air Force. |
| 1966 | Demonstrated accurate recognition of isolated digits. Invented technique to automatically identify talker. |
| 1967 )<br>1968 ) | Developed and delivered miniaturized voice controller for astronaut.<br>Developed NST to recognize digit strings for U.S. Post Office,<br>- operated in real-time in high noise environment<br>- for universal speech including many dialects, largest speaker population ever tested.<br>- spoken with no pause |
| 1969 | Invented technique to automatically identify language. |
| 1970 | Constructed and delivered speaker identification equipment to the Air Force and Army. Invented adaptive speech-recognition system. |
| 1971 | Developed programmable system for recognizing continuous speech utterances. |
| 1972 | Introduced commercial speech recognition system (VIP-100) for limited-vocabulary applications. |
| 1973 | The VIP-100 was selected by Industrial Research as one of the most significant new products of the year. |
| 1974 | Introduced direct voice programming of computer for NC tape preparation (VNC-100). |
| 1975 | Introduced a low-cost microprocessor-based voice data entry terminal (Threshold 500) to replace and/or complement intelligent terminal applications. It is ideally suited for large, multiterminal data entry systems. |
| 1976 | Introduced more sophisticated NC Tape Preparation system using voice programming (VNC-200). |
| 1977 )<br>)<br>)<br>)<br>) | Introduced user-programmable voice data entry terminal (Threshold 600) which is CRT/Teletype compatible.<br><br>Approximately 200 Threshold terminals are installed in various Government and industrial applications in 8 countries around the world. |

number of these recognition units are available and are used for experimental purposes. In addition, a number of Data General Nova 1200 and Nove 3 computers also are available for experimentation.

Several disk-based computer operating systems are also available for generating and debugging software. Some of these include 5 Megabyte disk storage and others 10 Megabyte storage. Paper tape reader punches and medium speech printers are also part of these disk oriented systems.

## REFERENCES


1. M.B. Herscher and R.B. Cox, "An Adaptive Isolated-Word Speech Recognition System", Proceedings 1972 Conference on Speech Communications and Processing.

2. T.B. Martin, "Acoustic Recognition of a Limited Vocabulary in Continuous Speech", Ph.D. Dissertation, University of Pennsylvania, May 1970.

3. T.B. Martin, "Practical Applications of Voice Input to Machines", Proceedings of the IEEE, Vol. 64, No. 4, April 1976.

4. M.B. Herscher and R.B. Cox, "Voice Programming of Numerically Controlled Machines", Proceedings 1977 International Conference on Acoustics, Speech and Signal Processings.

5. M.W. Grady and M.B. Herscher, "Advanced Speech Technology Applied to Problems of Air Traffic Control", NAECON '75 Record.

6. E.F. Grunza and S.G. Cohen, "A Voice Activated Control System for the Severely Handicapped", Proceedings 1977 International Conference on Acoustics, Speech and Signal Processing.

## BIOGRAPHICAL SKETCH

### Marvin B. Herscher

Mr. Herscher received the BSEE degree in 1953 and the MSEE degree in 1959 from Drexel University.

He was employed at RCA from 1953 through 1970 with the exception of 2 years spent as an officer in the Signal Corps at Ft. Monmouth, N.J. At RCA Advanced Technology Laboratories, he initially worked on applied research in the field of semiconductors. From 1960 to 1970 he was engaged in the field of pattern recognition and adaptive signal processing.

At RCA he was promoted to engineering leader in 1963 and Manager of Signal Processing in 1968. In these capacities, he was responsible for applied research in pattern recognition, adaptive logic and speech analysis and synthesis studies using feature-extraction techniques.

In 1970, Mr. Herscher became a cofounder of Threshold Technology Inc. (TTI) which was organized to develop commercial speech recognition equipment. He is co-inventor of the VIP-100 isolated-word recognition system which represents TTI's initial commercial product. Presently, he is executive vice president of Threshold, and is responsible for engineering, product planning, and R&D.

He has authored more than 30 technical papers, and is a co-author of the second edition of the Handbook of Semiconductor Electronics (McGraw Hill, 1962). He has been awarded six patents and has four additional patents pending.

## DISCUSSION

### Marvin B. Herscher

Q: Jared Wolf: You just mentioned a minute ago in the context of continuous speech recognition the adaptation to the user and I don't believe you do this at all in the isolated word systems and I wonder how come.

A: I don't believe I said adaptation to the user. I'm sorry if I did. What I said was basically for the system to be optimized for the user. You have to realize what Mike Grady was talking about, for example, is very similar to what we did ten years ago in the sense of looking at strings of events as detected by the acoustic detector and just looking for a sequential decisions. In order to accommodate large populations with very different dialects, for example, nan----versus nine, fo----versus four, ect., in various regions, you must have all the state diagrams that he showed expanded in very strange and complex ways. On the other hand if you limited recognition to a particular talker, you've got a much more restrictive and easier system to handle. When we did that, we threw away a lot of the extraneous paths that had to be accommodated in the general case. It was rather easy to do and worked quite well.

Q: Jared Wolf: Let me make my question a little bit clearer because I blew it the first time. By adaptation I meant, adaptation of the templets themselves. You occasionally talk about the need for retraining a templet or something like that and I'm talking about the adaptation of the templets in the same way that I think Dr. Plummer talked about this morning. In other words, tracking it over time for a user and keeping it up to date.

A: That's one of the things that might be a good topic for tomorrow. There are pros and cons in terms of doing it because depending on the application and how observant the individual is and what the starting accuracy is, it's very possible, for example, if the individual is not aware that he has made a mistake, that the system will start to diverge. As I said, there are pros and cons as to how you would handle it in an overall system. By the way, I forgot to mention one other thing too, --- the use of syntax itself is obviously very nice when you restrict the search space and everything works great except for one factor. Supposing the individual isn't completely familiar with the syntax and says words that are out of syntax, maybe legitimate words in the overall vocabulary but not within that little branch. In this case, the

probability is darn good that he is going to be forced into a misrecognition of one of the words that is in that branch. You've got to be very careful of things like that. So there are pros and cons in using syntax depending on the user and the overall system.

Q: Don Connolly, FAA: One of the things that I did but didn't report in the talk this morning but is in the paper, is that in training or getting a set of templets, (if you like, templets for my speakers), my data, and the results that I reported, include the learning function, whatever it is. Now I had a couple of speakers who never had to retrain any words at any time. I had a couple that had to retrain two or three words, four words, upwards of one or two times each. On the average each of the 10 or 11 speakers in each of the trials for each of the separate vocabularies retrained one word on the average, one word once.

A: Out of how many words, Don?

Q: Don Connolly: Well, out of 15, 20 depending on the number of words in the subset. Now, one of the things I did was save the last best set of templets in digital form on cassette and then I brought these people back in three months, six months and some of them nine months after they had last spoken to the machine. They hadn't even looked at it, they haven't even been near it and we got identical accuracies with the templets that were three, six nine months old.

A: That's one of the points that I mentioned but I didn't go into detail about it. I expect meybe we can talk about it tomorrow but in any of these systems, time stability is really important. Dr. Connolly has indicated that the particular feature set that we had used seems to work pretty well and holds up well with our recognition algorithms over long periods of time. True, you may have to re-train once in a while but once you reach a steady, reasonable templet the data is very consistent and holds up very well.

Q. Sam Viglione: You brought up a number of applications that address some of the problems that were discussed here and perhaps one of the things that we would like to know is the user acceptance and the performance within these environments. I would mention, for example, the stock exchange which appears to be a babbled type of environment with a lot of noise background similar to what's actually going into the system. How is the user accepting that environment and how does the system work? Another example is where you have the UPS environment, you have the RF link, you had physical exertion where actually the voice generating mechanism is being changed as the speaker is talking. How is the

user acceptance in that environment with the RF bouncing around inside of a metal room, too. How does the user accept the system and what is the performance in those kind of environments?

A. Let me talk about UPS first. We know when we started that job that the biggest problem that we would have would be the RF communications. Our part we could handle. We could do whatever we had to do to make the recognition work. The radio performance would be almost completely out of our hands. As it turned out when we got all done, - I won't give you the details because as I said I think you should pay for it yourself, - we ended up basically designing our own form of radio system. We actually have a special radio configuration designed to eliminate any of the null points or multi path that you're worried about. We sweated a lot of days on that one. The actual physical exertion of the operator turned out not to be too much of a problem because one of the things that you've got to do, in our system at least, is to overcome the long extensions of words caused by breath noise. Basically you're dealing with a very long exhalation of breath noise at the end of the word due to this exertion. The operators are panting, they're saying five-uhh, six-uhh. They're really working hard so you've got to have techniques to accurately detect where the true end of the word is versus where it appeared that the overall energy decreased. And you've got to be able to detect the true word ending; we do that by software. Fortunately, some of the features which are available in our system are present for speech but not for breath noise and we can use these features to handle that kind of a problem.

The environment at the Chicago exchange is extremely bad. The noise is very very bad and it goes up and down depending on how excited the brokers get. Opening is unbelievable - when the bell rings you can't hear yourself think. At closing, they go absolutely berserk and if you think its bad there, in New York the excitement is ten times worse. But basically we use noise cancelling microphones, and the system works reasonably well. There are some problems at Chicago mainly because the reporters are low level employees and are intimidated by the brokers who tell them not to talk to loud because they're interferring with their trades. Often, it ends up that they're whispering into the system, and it is really difficult to have a system operate well under those conditions. When they could talk using normal levels, by standing back a couple of feet from the brokers, we've had quite good success. We can run into problems when the excitement starts and the reporter gets intimidated and starts whispering.