PRACTICAL APPLICATIONS OF INTERACTIVE VOICE TECHNOLOGIES -- SOME ACCOMPLISHMENTS AND PROSPECTS

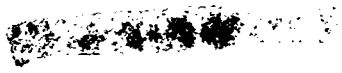
M. W. GRADY, M. B. HICKLIN, J. E. PORTER

LOGICON, INC.
TACTICAL AND TRAINING SYSTEMS DIVISION
SAN DIEGO, CALIFORNIA

INTRODUCTION

Logicon is a systems house, devoted to applying computers and electronics to bring new degrees of automation to complex systems. Logicon's efforts are characterized by the integrated application of advanced technology to products and services for industry and government. Although qualified in the various academic disciplines, Logicon's staff is primarily applications oriented, with a demonstrated record of accomplishment in the inventive and practical utilization of new technologies in complete user-based systems. Generally, then, Logicon is neither a research based organization nor an Original Equipment Manufacturer (OEM) supplier. Whenever currently available hardware can properly support an application, that hardware is utilized. Often the capabilities of various components are augmented through software enhancement and/or integration with other devices. In all cases, however, Logicon's paramount concern is with applications in turnkey systems.

This business philosophy is reflected in Logicon's interests in the advanced speech technologies; i.e., speech recognition and speech generation. Logicon's first association with the voice technologies was in 1969 when analog voice generation was utilized to automate a weapon systems trainer for the Naval Training Devices Center. Since that time, Logicon has continued to exploit the capabilities inherent in the advanced speech technologies. This is evidenced by noting that Logicon currently has (in-house) approximately 45 speech synthesizers and 15 speech recognition units that will be integrated into complete systems and delivered in the next several months. Logicon currently also has seven contracts with varied government agencies for programs utilizing speech recognition and/or speech generation. Each application is marked by the effective integration of the speech component into the total system. The voice capability is based on software enhancements of commercially available hardware chosen to reflect the specific requirements.



In a technological area receiving so much attention from research institutions and development laboratories, Logicon is proud of its record in the practical applications of interactive voice technologies. Logicon has developed systems which were only vaguely envisioned just a few years ago. Logicon is pleased to share some of these accomplishments with its colleagues through this forum, and to reflect on prospects for the future applications of automated speech technologies in real-time command and control systems.

ACCOMPLISHMENTS

The following paragraphs focus on three existing systems which typify Logicon's utilization of speech recognition and voice generation as interactive elements in complete turn-key systems. Note that in each of these systems, the voice technologies do more than simply enhance an existing man/machine interface. Rather, the technologies are employed as the cornerstones of totally new concepts made possible now with these automated speech capabilities. The potential for applications of this type seem all too easily overlooked. Developers of new technologies are often biased toward "proving" their technologies by demonstrating the direct substitution of the new for a well-established technique. If viewed as simply a technology replacement, it will be many years before speech recognition units will replace the more traditional, manual entry devices.

A more satisfying and justifiable approach is to consider replacement or automation of human tasks rather than replacement of some hardware equipment. Note that in each of the following systems described, the voice technologies are interactively combined with some measure of artificial intelligence to perform a task that would otherwise require the full attention of another person. The cost-effectiveness of such systems, especially when viewed over complete life-cycles, is not difficult to justify. In this way, automation, computers, and electronics combine most effectively to improve system productivity and to save money.

Flight Training Systems. As mentioned previously, Logicon's earliest exposure to the speech technologies was in 1969 when, under contract to the Naval Training Devices Center, we were involved in automating an experimental weapon systems trainer, TRADEC. This work was the precursor to today's highly successful Automated Adaptive Flight Training System (AFTS). The AFTS works in conjunction with existing flight simulators to automate the training syllabus associated with Instrument Flight Maneuvers (IFM), Ground Controlled Approach (GCA), Air-to-Air Intercepts (AAI), and Ground Attack Radar (GAR) operations. The AFTS has been developed and integrated into F-4E and TA-4J flight simulators.

Each of the AFTS modules incorporate the following design features:

a. Automated and adaptive flight training syllabi.

b. Standardized preprogrammed training scenarios.

c. Objective performance measurement and scoring.

d. Individualized, self-placed aircrew training.

e. Flexible and responsive instructor control.

f. "Strap-on" implementation - accomplished without modification to the basic simulator.

Both speech generation and recognition were to be utilized; although recognition was a relatively late entry. In the earliest phases, speech generation (Cognitronics and Metrolab voice drums) was used as the automated link between the computerized instructor and the trainee aircrew. GCA approaches were practiced by generating the appropriate advisories via the speech generation system. At the time (1969-1973), the voice drum was really the only technological device available. Fortunately, the GCA vocabulary was restrictive enough that these relatively limited-capability systems were wholly adequate. These systems and this application are of prime importance in terms of their historical significance. Speech technology was a basis for new concepts in (training) systems design.

In 1974, three separate factors came together to change the direction of voice generation in AFTS:

1. The voice drum technology was becoming increasingly expensive. Being an analog device, it was not sharing the benefits of the digital electronics explosion.

2. An electronic voice synthesizer, the Votrax® VS-6, was introduced which performed adequately.

3. AFTS application grew to include air-to-air intercepts, resulting in significantly increased vocabulary.

The enhanced AFTS consequently utilized the newly synthesized voice generation technology. It became literally true that it was no

---

® Registered Trademark.

more difficult to cause the computer to speak than to have the computer print (although one had to learn to "spell" all over again!). The system designers had complete flexibility in developing new vocabulary and hence new functions for the speech generation portion of AFTS. Perhaps most importantly, the enhanced AFTS demonstrated that operational flight crews could easily understand the synthesized speech even when engaged in a complex task such as a GCA or an AAI exercise. Synthesized speech had come of age, and was successfully demonstrated in a high fidelity simulation environment.

The AAI portion of AFTS, however, was lacking the full measure of automation. Very clumsy and artificial microphone-keying was required by the crew for AFTS to interpret where they were in the intercept. Specifically, the operational environment required certain specific actions on the part of the air controller (simulated by the AFTS) when the aircrew transmitted, for example, "contact", "judy", "lost contact." The AFTS required the crew to key their microphone to indicate these critical points. The results were clearly less than ideal. The solution to this problem became apparent - it was speech recognition. Speech recognition, however, had never been applied in the operational-like setting which exists in a high fidelity simulation environment such as the F-4 Weapons System Trainer (WST). A variety of critical questions were generated which could not be easily answered, including:

1. Will students undergoing AAI training conform to a standard phraseology for certain UHF transmissions, thus allowing recognition with usable accuracy?

2. How much training is needed to achieve usable accuracy levels? Training here refers to both machine training (that is, capturing the voice characteristics for each student that are later used during recognition), as well as student training or conditioning to use the acceptable (recognizable) phraseology.

3. Will the voice characteristics of the student drastically change under the simulated environment of an actual mission, thus affecting recognition?

4. Will the speech recognition hardware be able to reject the high levels of noise present in the WST audio system? Or will this noise mask the voice features critical to effective recognition?

A feasibility implementation study was initiated in 1975 to derive answers to these questions. The vocabulary consisted of 10 phrases for the pilot and 20 phrases for the weapons officer. Both

speakers utilized a single voice input preprocessor; reference patterns for both speakers were kept in core simultaneously. Significant integration problems occurred: e.g., the 400 Hz ac used in the cockpit for lighting, etc., interfered with the audio system causing large amounts of hum, noise and distortion. This made the feature extraction process less reliable than had been experienced in the more controlled environments of a laboratory or other setting. This problem was largely solved by careful filtering and shielding the audio signal.

User acceptance also presented a challenge. Because the verbal behavior of the aircrew is a relatively insignificant element of their primary function, the users resisted conforming to the "approved" vocabulary, and configuring the system with their voice characteristics. (This observation was in direct contrast to experience with controller training where the student's vocal procedures are critical to his mission. Refer to the following subsection.)

Despite these difficulties, the AFTS experiments with speech recognition were clearly a success. The training system is significantly enhanced by the automated pseudo-instructor and pseudo-controllers. It truly is exciting to witness the real dialog between man and machine that can occur in AFTS with speech recognition and voice generation. The following example typifies this intercourse of a truly interactive voice system:

AFTS:	"Phantom 1, cleared for reattack"
Aircrew:	"Say again"
AFTS:	"Phantom 1, cleared for reattack"
Aircrew:	"Roger"
Aircrew:	"Phantom 1, Contact"
AFTS:	"Roger, contact is target"
Aircrew:	"Phantom 1, Judy"
AFTS:	"Roger, Judy"
Aircrew:	"Phantom 1, Lost Contact"
AFTS:	"Phantom 1, you have a target at ---"
Aircrew:	"Phantom 1, Roger"

Controller Training Systems. Based on the successes of the early automated and adaptive flight training programs, in 1972 the Naval Training Equipment Center sponsored Logicon in an investigation of similar teaching concepts applied to controller training systems. The foundation of any automated adaptive training system is the ability to monitor the relevant behaviors of the trainee while he is performing his

tasks. In pilot training, the trainee's interactions with the simulated aircraft controls is monitored. In controller training, however, the verbal behavior of the trainee must be monitored. The emergence of computer-based speech recognition was therefore welcomed as potentially providing the basic technology with which automated controller training could be realized.

The Ground Controlled Approach Controller Training System (GCA-CTS) subsequently became (and remains) Logicon's crowning achievement in the application of interactive voice technology in training systems. The first system delivery of the GCA-CTS more than three years ago, represented the first application of automated speech recognition to a sophisticated training problem.

Subsequent deliveries of the GCA-CTS included speech generation capabilities and a variety of improvements in the training methodologies. It is important to note that the GCA-CTS demonstrates the total integration of the speech technologies into the whole system. Speech synthesis is used to prompt beginning students in learning the correct GCA phraseology. Moreover, the synthesizer verbally instructs the student during replay, describing the errors committed by the student. Speech recognition is used to effect changes in the movement of the simulated aircraft, and to provide the inputs to the performance measurement subsystem. The speech understanding unit, therefore, replaces a "pseudo-pilot" and, at the same time, allows automated and adaptive training.

The limitations of the recognition technology (e.g., requirement for a priori reference data) present no difficulty because they are smoothly incorporated into the total training program. (The student is learning the vocabulary at the same time as the computer is developing reference data.) Because the vocal behavior of the student is critical to his task, he is a willing and cooperative participant. Minimal unnatural speech stylizations are readily accepted and generally easily learned. These observations point to some important lessons to be learned about the application of this new technology: the speech capabilities must be totally integrated into the man/machine environment, and the benefits available must be clear to the user.

Operational Systems. The Automated Command Response Verification (ACRV) System represents an application of the voice technologies to an operational versus training problem. Again, the basis of the ACRV concept demands viable speech recognition and generation capabilities.

ACRV was conceived as a potential aid to the verbal communications link between a ship's pilot or conning officer and the helmsmen. The system recognizes the commands given the conning officer, and at the

same time, monitors the various ship control surfaces. These two sources of information are then compared in a computer. When a mismatch (or error condition) is detected, the system issues a verbal advisory, warning bridge personnel of the potential problem. The ACRV system is totally passive, in that no action is ever initiated by the system. The system does not, for example, ever issue a command; to do so, would not only usurp the authority of the conning officer, but would add to confusion on the bridge in times of stress.

Under contract to the Department of Transportation (DOT), Logicon developed an ACRV demonstration system in its engineering laboratory to establish the technical feasibility of this concept. The model utilized a specially constructed ship control console (helm and engine controls; rudder, RPM, and heading indicators) as well as speech recognition and synthesis equipment.

The ACRV system provides a convincing demonstration that the concept of applying the automated speech technologies to a safety application is indeed technically feasible. The most demanding vocabulary set was chosen to demonstrate that even subtle differences in long phrases could be distinguished. This large vocabulary demanded a great deal of ACRV software to perform the understanding of a large and diverse set of commands in order to behave in an intelligent fashion. The ACRV system demonstrates that automatic warning systems need no longer be conceived of as merely attention-getting alarms associated with specific error conditions (as is provided by aircraft stall warnings); but rather the ACRV is one system which is able to distinguish a wide variety of errors and, furthermore, is able to provide an exact report of the error just as a crewmember might. Thus, attention is called to the error condition which can be corrected before danger threatens.

CURRENT APPLICATIONS

Logicon is continuing to enhance the systems described in the preceding section. The Logicon-AFTS (with both recognition and generation) is in production and has been acquired by the U.S. Air Force for 16 F-4E simulators throughout the world. Additional AFTS systems are being developed for other aircraft, such as the A-7A. The laboratory GCA-CTS has sufficiently evolved so that a self-standing, experimental prototype GCA-CTS is being developed for evaluation at the Navy's Air Traffic Control School. The next step in the ACRV development cycle is an assessment of operational acceptability using a shiphandling simulator and experienced conning officers.

Other programs are underway at Logicon which also utilize the interactive voice technologies in real-time command and control systems. These programs currently include:

a. Landing Signal Officer Training - An automated adaptive training system for the LSO is under study. Important elements of the envisioned training system will be speech recognition and generation.

b. Air Intercept Controller (AIC) Training - The AIC vocabulary is significantly more complex than the GCA or LSO vocabularies. The automated AIC training problem thus represents a significant advance in the application of the speech technologies to training systems design.

c. Pseudo-Pilot Replacement - Many complex training systems utilize console operators to interpret student commands and to enter data into the simulation computers. This task is accomplished via speech recognition in the GCA-CTS; the applicability of using this technology in other training environments is being pursued.

d. Pseudo-Instructor Functions - Both the AFTS and GCA-CTS utilize the speech technologies to simulate many functions normally performed by the instructor. This concept is being expanded in the development of instructional systems for the B-52 and KC-135 simulation training systems; and the Instructor Support System (ISS) for the F-14 flight trainers.

e. Cockpit Design Studies - Working with a major airframe manufacturer, Logicon is involved in the study of utilizing interactive voice technologies in cockpits of future (1985+) aircraft. The impact of these technologies on in-flight performance measurement and crew training also is being assessed.

HARDWARE AND SOFTWARE COMPONENTS

Each system which has been described in this presentation utilizes an isolated word or phrase recognition capability and/or a synthesized speech capability. This section describes in greater detail specific technical aspects of these system components. It is important to observe that Logicon has no formal commitments to any hardware manufacturer, exclusive of the usual OEM agreements. Equipment is chosen solely on the basis of capability (vis-a-vis the intended application) and cost. Various other speech-based system components have been formally and informally reviewed by Logicon and many are ideal for applications other than those described herein. Logicon does not intend to endorse any particular manufacturer in this review.

Speech Generation. Logicon has utilized electronic voice synthesizers, specifically the Votrax® VS-6, since 1974. Except where naturalness is a firm requirement, the Votrax® has demonstrated completely acceptable voice quality. Vocabulary flexibility and low cost are particularly attractive features of this synthesizer.

A variety of software tools have been developed at Logicon to support the development of speech-generation-based systems. A peripheral device driver has been written for the Data General Corporation operating systems, for example, which enables the user to communicate to the speech synthesizer in meaningful ASCII phoneme strings through standard system calls, just as if one were communicating with a teletypewriter through ASCII word strings. This capability significantly eases the conversion of new vocabularies to inflection/phoneme commands, since the synthesizer is available to the standard text editor. A phrase composition program also has been written to enable users to construct new phrases for speech output using vocabulary words previously converted to phoneme commands.

Speech Recognition. Logicon has utilized voice input preprocessors developed by Threshold Technology, Inc. (TTI), since 1973. These preprocessors sample the speech approximately 500 times per second and detect the presence or absence of some 30 speech features. This information is relayed to the computer where the software (described in the ensuing paragraphs) performs the recognition algorithms. TTI preprocessors have been chosen for each application to date strictly on the basis of performance (the unit appears to be a nearly deterministic sound classifier), flexibility (vocabulary size, phrase length, etc., are software, not hardware, limitations), and cost (no expensive array processors or dedicated computers are needed to support the recognition process).

The isolated phrase recognition software utilized by Logicon is based on the algorithms developed by Threshold Technology. Significant enhancements and extensions to TTI's approach have been adopted however. These include:

a. Long phrases (2-3 seconds) are recognized with high accuracy. Reference patterns are 1024 bits vice 512 bits.

b. Effective schemes have been developed for distinguishing between the small differences that often occur in phrases of the vocabulary (e.g., "slightly above glidepath" and "slightly below glidepath").

® Registered Trademark

c. Rapid-fire voicings (several phrases, each separated by less than a half-second) can be accommodated.

d. A digit extraction algorithm has been developed for recognizing the final digit in a long phrase with high accuracy.

e. Effective use is made of the level of confidence in the recognition process. The system thus is often able to distinguish between user errors and machine (recognition) errors.

Most significantly, perhaps, the entire speech recognition software subsystem is packaged as a FORTRAN compatible module executing under Data General Corporation's Real-time Disk Operating System (RDOS). This package enables the almost immediate integration of a speech recognition capability into any FORTRAN-based RDOS program. To minimize core requirements, all the reference patterns are stored on the disk and selectively retrieved in real time when they are needed. Some very clever software structures permit this dynamic data swapping and still provide quick recognition of spoken commands. Another benefit is that, using this scheme, the vocabulary size is limited only by more practical considerations, such as training time, etc. The scheme would be especially useful in highly structured vocabularies since this would further limit the amount of data which must be retrieved from mass storage.

Based on the success in Logicon's speech application programs, other tasks have been identified as amenable to an interactive voice-based automated system. In addition, the problem of training the user in correct pronunciation and in use of the radio terminology, operational brevity codes, "standard commands", etc., is itself a subject for study. Finally, experience with speech recognition has highlighted certain risk areas associated with the recognition of some phrases. Identification of these problem areas early in a system's development cycle is central to finding effective solutions.

Aware of each of these requirements, Logicon has developed a highly flexible development tool called the Voice Data Collection (VDC) program. The VDC program provides the framework around which the system designer - at the user level - can:

a. Define vocabulary phrases associated with essentially any application.

b. Preprogram the presentation of phrases or groups of phrases to the speaker via text and/or computer-synthesized sppech, hence resulting in an effective environment in which the vocabulary phrases can be learned, and in which the

fidelity of reference patterns extracted during this learning phase can be enhanced.

c. Test the ability of existing hardware/software algorithms to recognize these phrases, and extract hardcopy on recognition reliability and potential system confusions.

Performance and Lessons Learned. One of the most critical elements of the total system, vis-a-vis good recognition rates, is the methodology associated with capturing the voice patterns of potential users. Logicon's experience has pointed to the importance of extracting voice characteristics in a fashion which replicates as nearly as possible the environment (ambient noise, stress, etc.) in which recognition will later be required. An interactive system which fluctuates between "training" and "validation" is highly beneficial. Users unconsciously (presumably) modify their speaking style to effect good recognition. In general, the longer one has used the system in a direct validation feedback mode, the better is his recognition rates. (There is a phenomenon of learning to talk to the box!)

The ACRV application described earlier was supported by a recognition capability encompassing 64 words or phrases. The vocabulary list was considered subjectively difficult since many phrases were syntactically similar. For users unfamiliar with both the vocabulary and the speech systems, approximately two hours of voice data collection and validation were required to achieve consistent accuracy in the 94 percent to 98 percent range.

FUTURE DIRECTIONS: LIMITED CONTINUOUS SPEECH RECOGNITION (LCSR)

Automatic speech recognition has been shown to offer opportunities for significantly improving the efficiency and effectiveness of training systems. Systems developed, and in operation, which demonstrate this practical benefit of speech recognition in training systems include the GCA-CTS and the AFTS. On the basis of experience gained in these systems, it is clearly desirable and appropriate to expand the use of automatic speech recognition in training systems.

Many training applications can be supported adequately by a capability to recognize isolated words or word groups automatically; the aforementioned applications are of this type. However, in some applications isolated word recognition is not adequate. An automated training system for training air intercept controllers, for example, requires recognition of numerical data, naturally spoken as an unbroken sequence of digits. In this and similar applications, the number of digit sequences of interest precludes the use of isolated word recognition

algorithms via the artifice of treating each possible sequence as a potential explanation of an utterance.

Under contract to the Naval Training Equipment Center, Logicon has been investigating LCSR during the past year. A novel approach toward solving the LCSR problem was conceived by Logicon in 1976. Based on concepts from the theory of mathematical machines, a simple sequential recognition procedure was modeled as a finite automaton. Continuous speech, it was postulated, could be characterized and recognized on the basis of observing:

a. The characteristic classes of output from a preprocessor.

b. The order in which these occur.

c. The characteristic time durations between the output samples.

The method of discovering the characteristic output classes and time durations is a direct automated examination of speech data. An initial implementation effort was defined to determine if indeed these assumptions were valid for the 10 digits and the word "point". A Threshold Technology preprocessor and Nova minicomputer presently support the research.

Experience in applying automatic speech recognition to practical training systems has revealed several special characteristics of the LCSR problems which arise in this class of systems. These special characteristics made the training LCSR problem much more specific than what is generally referred to as the "limited continuous recognition problem" in the technical literature.

Logicon is convinced that it is essential to scope tremendously complex problems, such as connected word recognition, to both focus the attention of industry and also to increase the probability of success by developing the most limited capability consistent with the system requirements. Several features which localize the training LCSR problem within the larger domain reported in the literature are discussed below. While not all of these characteristics are universally shared by all LCSR problems arising in training applications, it is true that any solution to the LCSR problem compatible with these characteristics would meet the requirements in most training applications.

a. A small vocabulary is involved. Many training problems entail vocabularies of 20 words or less, and often recognition of fewer words would be a useful capability. The 10 digits in combination with a few control words is a fairly representative and common case. Using a mixed

strategy of isolated and continuous speech recognition techniques can sometimes reduce the required vocabulary size of the continuous part of the problem even further.

b. The vocabulary is fixed. Within a given training application, the vocabulary changes with a half-life measured in months or years. As a result, rapid accommodation of vocabulary changes, while attractive, is not an important requirement. Techniques which entail detailed, and perhaps time-consuming, off-line analysis of the vocabulary items are therefore of no particular disadvantage.

c. Semantic, syntactic, and other higher knowledge sources are often nearly or completely irrelevant. This observation is typified by the numerical data entry problem, where strings of digits must be recognized, with essentially no hard data available in the remainder of the system which can be used to predict what the spoken digit string might be. In many cases, a priori probabilities can be assigned to gross features of the utterance, such as the number of digits in the utterance, or the identity of the first digit. Within the utterance (i.e., for non-initial words) it often occurs that the branching factor is essentially equal to the size of the vocabulary. The fact that a training system has to deal specifically with errors committed by the trainee exacerbates the problem, as deviations from proper syntax, for example, may be both more likely to occur and more interesting in themselves in the training environment than in the operational environment.

d. Real-time operation is necessary. Effective training often requires very quick response to trainee vocalization, either to preserve realism of a simulated environment or to minimize the latency between responses and reinforcement. A time lag of less than 2 seconds between completion of an utterance and recognition is often required.

e. Recognition accuracy must be high. Trainee motivation, and thus training effectiveness, drops precipitously with any decrease in a training system's reliability, and recognition failures are perceived as just another variety of system failure by the system user. The supposition that low recognition accuracy can be tolerated in training systems is often supported by the argument that the purpose of the system is to teach correct verbal behavior; and hence, the careful enunciation required for good recognition can be demanded of the trainee. This argument is fallacious for two reasons:

1. Few training systems have precise enunciation as an important training objective.

2. Within the present state-of-the-art, recognition accuracy in the high 90 percent region is only attainable with audio input which is very understandable to the human ear; careless enunciation significantly degrades the already less-than-perfect recognition accuracy currently attainable.

f. Speaker independence, while convenient, is not a necessity. Training systems which warrant a dedicated speech recognition capability tend to be associated with tasks which require several hours or more of training. A small amount of time spent adapting the system to the trainee's voice is rarely a significant drawback, particularly since this adaption period can sometimes be treated as part of the training experience wherein the trainee learns the vocabulary or how to operate the training system.

g. The computational requirements should be compatible with central processors on the scale of mini-computers or even smaller systems. This is simply an empirical observation on the economics of training systems. The computers used in training systems tend to be dedicated, and the training systems tend to be of such a scale and have development budgets which can accommodate the cost of mini or micro-computers, but often not the cost of a large main-frame. Counter examples can undoubtedly be found, but experience indicates they are the exception rather than the rule. This same observation applies to special-purpose hardware which supports the front-end analysis of the analog speech signal. Sophisticated, special-purpose preprocessing hardware can become very expensive and hence it is desirable to utilize established, commercially available components if possible.

The technical literature reveals some trends in continuous speech recognition which can be interpreted as augering well for the line of inquiry being pursued. Some of these trends are discussed below.

There is a trend toward de-emphasis of segmentation into classical phonemes and specific phoneme recognition. Earlier efforts focused on recognizing speech phoneme-by-phoneme, with articles appearing on the difficulties of recognizing particular phonemes. The tendency now is to

treat the preprocessor more nearly as a sound classifier, and to ignore preconceived notions of what the speech data received from the preprocessor are like. The reason for the tendency is that reliable segmentation into phonemes turned out to be impossible, dispelling the early hopes that the internal reference representations of words could be some simple variation of familiar phonetic spellings, modified by phonological rules.

It follows from the failure of rigidly phoneme-oriented recognition that there is a tendency to go to the speech data (that is, develop algorithms for processing real speech data) to determine its recognizable characteristics. This is in contrast to the early reliance on the obvious phonemic content of words to be recognized. The present recognition techniques being developed therefore tend to have two parts, first the recognition technique per se; and second, the techniques for deriving relevant parameters (such as Markov transition probabilities or likelihood-measure thresholds) from large samples of speech. This trend marks the demise of the early influence of linguists and phoneticians on speech recognition research.

There is also a recent trend toward sequential decoding of the speech signal instead of exhaustive hypothesize-and-test recognition methods. The distinction between these two approaches becomes blurred as the methods for optimizing the search of the test space become more and more efficient. Interestingly, both HARPY and Martin's early efforts are essentially sequential in nature. Both use a transition state model to determine a limited set of next-possible features. In the case of HARPY, this was a considerable simplification over its predecessor's models, which entailed probabilities of transitions to each of a large set of possible next states.

The approach adopted for the LCSR effort being conducted by Logicon conforms to each of the trends mentioned above; namely toward:

a. Treating the preprocessor as a sound classifier.

b. Emphasizing the derivation of the recognizable speech characteristics from real speech data.

c. Sequential decoding.

The investigation began by collecting a large number of utterances from a single speaker. The utterances were carefully chosen to observe the speech data in the presence of varied contextual influences. Nine-hundred-ninety utterances were recorded for a total of 3150 words. These data were divided into a training set, an interim test set, and a test set. The training data were further divided into example spaces for each vocabulary item.

A class of sets of sounds output by the chosen preprocessor was defined. Borrowing some terminology from the theory of formal languages, the sounds input from the preprocessor are called letters. The characterizing sets of sounds postulated by this approach are termed transition letter sets. An heuristic algorithm for finding the transition letter sets, and their order, was used to search the example spaces containing each vocabulary item. A remarkable amount of structure was found indicating that there are invariant structural features in the speech data which are reliable enough for use as a basis for recognition.

Having distinguished the sound groups which reliably occur in samples of each vocabulary item, attention was focused on the residual sound groups in the speech data. These data, termed loop letter sets, were demonstrated to be potentially effective in reducing the number of false recognitions. A computer program was implemented for finding the smallest collection of loop letter sets which accommodate the example spaces. Surprisingly, the resulting residual sounds were found to occur infrequently, indicating that the transition letter sets contain most of the sounds which comprise the entire word.

The collections of transition and loop letter sets for each vocabulary item were exercised over the interim test data. Statistical models were developed to describe the observations associated with:

a. The time durations in which the machines dwelt in each transition and loop state.

b. The violations of the transition and loop letter sets which prevented machines from continuing through to completion when bona-fide vocabulary items were actually spoken.

c. The occurrence of artifacts; i.e., machines erroneously going to completion.

d. The time-based overlaps and gaps associated with multiple machines running simultaneously over connected speech.

These statistical models were incorporated into the design of the final Machine Execution (MEX) algorithm and Machine Interaction (MINT) algorithm. Implementation of these algorithms in the computer program LISTEN (Logicon's Initial System for the Timely Extraction of Numbers) is currently in progress. Initial recognition accuracy estimates hopefully will be available before the end of this calendar year. Although LISTEN is being coded in FORTRAN, Logicon expects nearly real-time time operation on a Data General minicomputer.

The LCSR capability being investigated by Logicon is specifically tailored to the unique requirements of connected word recognition in training systems design. Again, Logicon's approach is oriented toward supporting a practical and immediate application area; namely, an automated training system for air intercept controllers. If these efforts are successful, clearly the application of automated speech recognition will advance into new areas presently not supported by isolated word systems.

BIOGRAPHICAL SKETCHES

MICHAEL W. GRADY is the Technical Manager at Logicon's Tactical and Training Systems Division for programs utilizing the advanced speech technologies. He brings to this position several years of experience in real-time systems design and development, particularly in both large and small scale training programs. Mr. Grady received the Master of Science degree from the University of California in 1969.

MARY B. HICKLIN is currently the Project Leader at Logicon for the Ground Controlled Approach - Controller Training System. Ms. Hicklin has been intimately involved in all of Logicon's voice-related projects since 1973. She has also been responsible for the definition and development of performance measurement subsystems in various training systems. Ms. Hicklin holds a Bachelor of Science degree, conferred by San Diego State University in 1969.

JACK E. PORTER is the Senior Analyst at Logicon's Tactical and Training Systems Division involved with requirements analysis and the application of mathematics to systems analysis and engineering problems. He is the principal contributor in the effort to develop a Limited Continuous Speech Recognition capability. Mr. Porter holds a Master of Arts degree in Mathematics from the University of California (1974).

DISCUSSION

Michael W. Grady

Q: Leon Ferber: How do you configure the LCSR system for the speakers voice?

A: That problem will really be addressed in what I hope will be the next phase of this program. One of the limitations that we made on our system for the time being was to totally ignore the "training" problem. I sat a total of about six hours developing programs and working on it since then. But clearly we must yet find more acceptable configuration method.

Q: Bob Plummer: If you don't have 11 parallel processors to do the word spotting, how do you time share between those at the early stage of the front end?

A: We simply sequence through them in serial fashion. Luckily the procession could be done in FORTRAN and still be handled in real time.

Q: Jared Wolf: I would like to take an exception to something not that you said but something that you wrote in your paper. You seem to constantly predict the demise of phone oriented recognition and you point with confidence to your approach. Apparently the paper was written before you had something to be confident in and to the HARPY system but I just wonder if you are really being serious there?

A: Jack Porter: I take the blame there entirely. I wrote those words approximately a year ago based essentially on the perception that linear predictive coding seemed to be of tremendous interest to speech researchers in recognition area. It appears to me that when you use something like LPC coefficients or the residual you're not taking any consideration whatsoever in the phonetic significance of the underlying sounds. I would like to withdraw that statement. It's premature and is based on inadequate data. Apologies given.