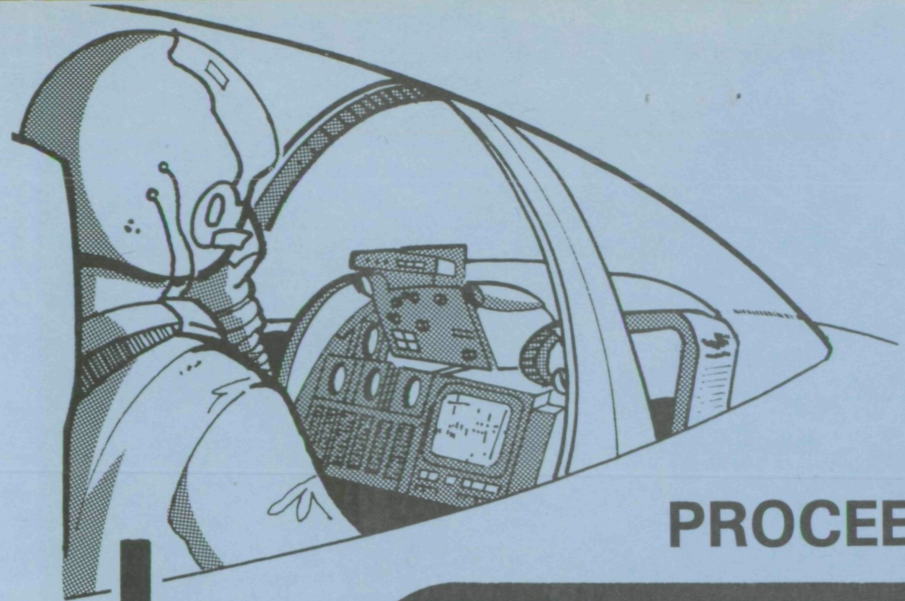
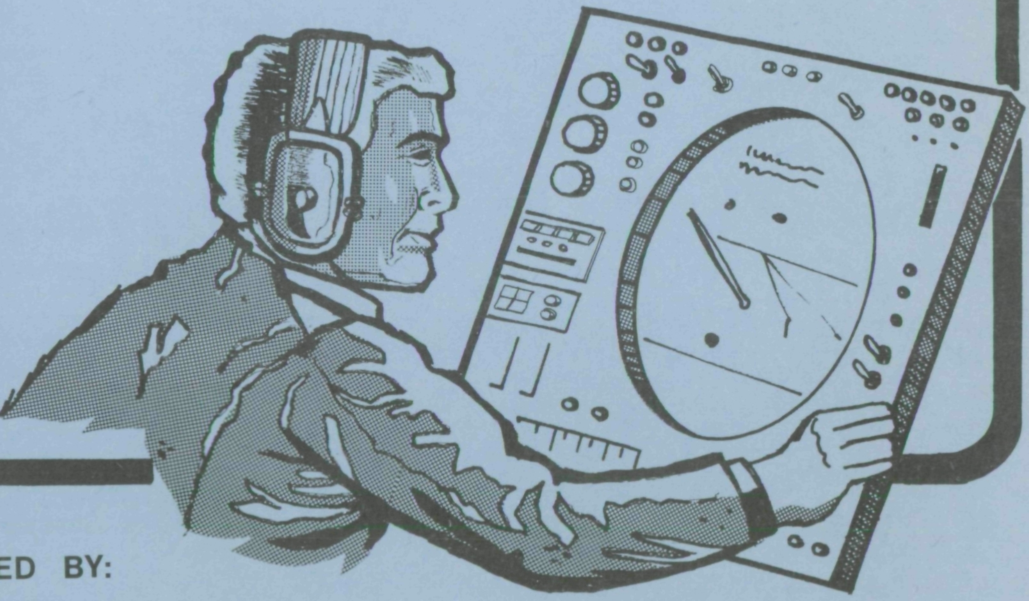
CR-152283



# PROCEEDINGS

## VOICE TECHNOLOGY FOR INTERACTIVE REAL-TIME COMMAND/CONTROL SYSTEMS APPLICATION



SPONSORED BY:

NASA
National Aeronautics and
Space Administration

**Ames Research Center**
Moffett Field, California 94035
**6 - 8 DEC. 1977**

PROCEEDINGS

# VOICE TECHNOLOGY FOR INTERACTIVE REAL-TIME COMMAND/CONTROL SYSTEMS APPLICATION

December 6-8, 1977

NASA, AMES RESEARCH CENTER
MOFFETT FIELD, CALIFORNIA

SPONSORED JOINTLY BY THE

NAVAL TRAINING EQUIPMENT CENTER
Orlando, Florida

NAVAL AIR DEVELOPMENT CENTER
Warminster, Pennsylvania

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
Ames Research Center, Moffett Field, California

Co-Chairmen and Editors:

Dr. Robert Breaux, Naval Training Equipment Center,
Orlando, Florida

CDR Mike Curran, Naval Air Development Center,
Warminster, Pennsylvania

Dr. Edward M. Huff, NASA, Ames Research Center,
Moffett Field, California

## TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

(This page intentionally left blank)

OMIT TO P.10

# CALL TO ORDER

CDR P. M. CURRAN[1]

NAVAL AIR DEVELOPMENT CENTER
WARMINSTER, PENNSYLVANIA

This symposium and workshop on Voice Technology for Interactive Real-Time Command/Control Systems Applications is sponsored by three Government agencies. Our co-chairman, Dr. Robert Breaux represents the Naval Training Equipment Center, Orlando, Florida, and co-chairman, Dr. Edward Huff, represents NASA, Ames Research Center. I am CDR Mike Curran, your co-chairman representing the Naval Air Development Center, Warminister, Pennsylvania.

The genesis for this symposium and workshop is three-fold. The Speech Understanding Workshop sponsored by the Defense Advanced Research Projects Agency (ARPA) in November 1975 closed with an interchange of ideas among the participating researchers and managers. All agreed that follow-on DOD meetings,or Government meetings, or public sector meetings, would be of value.

The second driving force is that as personnel involved in Voice Technology contacted each other during the last several years, they repeatedly stated a desire to participate in some forum to exchange information concerning current developments in Voice Technology. All managers and researchers I have been in contact with have expressed their inability to keep abreast of the work of all identified players.

The last inpetus dates back to over a year ago when CDR Paul Chatelier, Naval Air Systems Command Program Manager for Human Factors Engineering, saw a need for an exchange of information concerning Government Voice Technology efforts. As a first step, representatives of NADC, NTEC, and NASA met and discussed their programs. The next step was that CDR Chatelier provided the funding for this symposium and workshop which includes both Government and Industry participants. Finally, without the close cooperation of the co-chairmen and the generous support provided by Mr. Jim Duva of NTEC, and our host, NASA, Ames Research Center, this meeting would not have occurred.

Now to mention a few items concerning the mechanics of our meeting--I beg your indulgence, but no smoking or drinking in this room or building. Our morning and afternoon coffee breaks will be located in the foyer of Building 200, which can be reached through the walkway adjoining this building. We would like you to confirm your participation in the Life Sciences Tour for Thursday afternoon by checking the list on the table outside this room. For your convenience, a pay telephone is located

[1] The opinions expressed here are those of the author and do not necessarily reflect the official policy of the United States Navy.

down the stairs outside. Those who need Government Orders stamped, please place them in the box on the table outside and they will be available for pick-up before the workshop begins on Thursday.

If you have any questions concerning physical arrangements, please feel free to see Nancy Frazier from Telcom Systems, Inc., who is serving as the coordinator for this gathering.

When you have a moment, you might review the General Information page in your Workbook.

Now, I would like to introduce Dr. Syvertson, the Acting Director of NASA, Ames Research Center, who will present the Welcoming Message.

Dr. Syvertson--

## WELCOMING MESSAGE

MR. SYVERTSON

NASA, AMES RESEARCH CENTER
MOFFETT FIELD, CALIFORNIA

Thank you, Mike. I'll be relatively brief. I want to welcome everybody here on behalf of the Ames Research Center. We're very pleased to host this symposium, especially with the wide variety of organizations represented. In looking over your agenda, and also in talking to some of the people who helped organize this meeting, I was interested to see how rapidly this technology is advancing, and also to see the wide variety of applications which are being considered. We've had a program here at Ames, not a particularly large program, but an active one, and we've been aware of the potential applications to aircraft, especially the possibility that automatic speech recognition can reduce pilot workload. This reduction is getting to be an important consideration as modern aircraft become more and more complex.

That, however, was the only part of the activity I was aware of and I am pleased and surprised to see the other aspects. The Center is very interested in the field. We hope that you will enjoy this conference. Dr. Ed Huff from Ames, and anybody else on our staff will do everything they can to help make your stay enjoyable. I am sure they will be happy to do that, so just ask and we will try and take care of you. I do want to welcome you and say again that we are very happy to host this meeting.

(This page intentionally left blank)

## SYMPOSIUM/WORKSHOP OBJECTIVES

CDR MIKE CURRAN, Ph.D.[1]

NAVAL AIR DEVELOPMENT CENTER
WARMINSTER, PENNSYLVANIA

R&D managers wish to direct programs which are responsive to well defined and clearly specified sets of requirements. All too often R&D programs and symposia lack such definition and specificity for their requirements. For symposia, the end result is often only a set of proceedings which document papers presented and related discussion. Hopefully, to avoid these pitfalls, I will attempt to specify the several purposes or goals of this sympsium and workshop. When the three days of meetings end, I shall attempt to make some judgment about how adequately we met the state goals.

One obvious purpose for our gathering is the need for an exchange of information since voice interactive systems represent a booming technology area. Currently, more than a dozen government agencies are engaged in supporting voice development efforts. Inspite of the termination of the ARPA SUR Program, the overall picture is one of increased government and industry IR&D funding. Admittedly, there is an apparent shift from basic Research and Exploratory Development dollars to Advanced and Engineering Development dollars. The number of current industry players reflects a dramatic increase in interest and attention to this technology area. A growing number of companies are designing and packaging voice recognition units to meet the needs of very specific markets. An increasing number of firms have the resident capability and expertise to accomplish the implementation of interactive voice systems for specified system applications. In brief, more requirements for system applications; more viable programs; more players; and more dollars comprise the picture for today and the near future. Unless we exercise every occasion to exchange information, opportunities may be lost for this technology area to maximally impact and effect both identified and potential areas of application.

The goal of this symposium is not simply to provide another forum for information exchange concerning R&D for voice technology areas. While such exchanges are always useful, meetings of the IEEE and other groups provide opportunities to appraise both managers and researchers alike of technical advancement and progress in voice technology areas. Our purpose is to exchange information specific to the application of voice technology to interactive real-time command and control systems.

[1] The opinions expressed here are those of the author and do not necessarily reflect the official policy of the United States Navy.

The approach of this symposium is to first present an overview of the ARPA SUR Program objectives and approach. Next, a limited number of research efforts originally identified with, and funded by, the SUR Program will be presented. The goal of these presentations is not to explicate in detail the capabilities of large-scale systems developed under ARPA sponsorship. Rather, the goal is to look to these long-term development efforts to see if we can draw upon a wealth of accumulated experience to identify technology gaps and voids which remain to be addressed if successful applications are to be effected. However, these various approaches must be sufficiently understood and appreciated to see how they can strengthen and contribute to presently identified government requirements for system applications.

Next, a number of government R&D efforts will be presented. Requirements for system applications will vary according to the goals and mission of the specific organization. Some programs will reflect a long-term development effort, while others will have limited specific goals. All the government programs reflect a need to incorporate both the achievements of the ARPA SUR Program, and the gains made by industry IR&D programs. We have asked each government participant to spell out his program in terms of pre FY-78 efforts and post FY-77 efforts. In addition, he has been asked to anticipate the near, mid, and long-term technology requirements necessary to support his rpogram. It is hoped that we can gain a comprehensive picture, across government agencies, of voice technology R&D requirements. Such information will help government sponsors avoid supporting redundant efforts, and identify areas of common interest and concern which can benefit by the conduct of collaborative programs. Government agencies should be able to evolve mutually supporting programs which allow for a timely resolution of identified technology gaps and voids. Industry should benefit by gaining an awareness of our needs. Hopefully, they will direct their endeavors to be responsive to our stated requirements.

The last part of the symposium will present a number of industry IR&D efforts which have demonstrated promise for advancing the state of the art of voice technology. Specifically, we are interested in industry efforts which enhance our capability to achieve real-time interactive command and control systems applications.

The workshop will address issues and problems identified by the various presentors. All presentors have been privy to the prepared papers which will appear in the proceedings of this symposium. In addition, the three co-chairman will be noting points of concern raised furing the discussion period following the presentation of each paper. It is hoped that not only will the workshop identify salient issues, problems, and proposed solutions, but that it will serve as the occasion to formalize the initiation of a permanent vehicle to exchange information concerning voice technology systems applications. requirements, and developments.

Hopefully, the goals for the symposium and workshop are clear. Surely, they are ambitious. If we meet them, we will share a satisfaction similar to that experienced by a vintner as he uncorks a bottle of promising wine. After a new wine, or technology area, has sufficiently matured or aged, we wish to test the fruit of our labors. With a new wine the test is whether our palate is pleased. With an emerging technology area, the test is whether we achieve successful applications. In both cases, the fear is that once the effort we have labored with so long has surfaced, it may prove wanting, and we may find that our test was premature. However, for the area of voice technology I am optimistic. I believe that it has sufficiently matured, and that it will not be found wanting when it is applied to meet requiremtnts for system application--if we correctly utilize the presently available products of this technology area, and if we also direct our future efforts to obtain those products required to support identified user requirements.

(This page intentionally left blank)

## DARPA OVERVIEW

LT. COL. DAVID CARLSTROM

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
ARLINGTON, VIRGINIA

Lt. Colonel Carlstrom gave a brief overview of the DARPA Speech Understanding Research program. He identified the main system contractors as Bolt, Beranek & Newman, Carnegie-Mellon University, Lincoln Laboratory, System Development Corporation and SRI International. The specialist contractors were Haskins Laboratory, Speech Communications Research Laboratory, Univac, and the University of California at Berkeley.

Lt. Colonel Carlstrom reviewed the research objectives of the DARPA effort. He recommended a document by Newell, et al titled Speech Understanding Systems: Final Report of a Study Group for those interested in a detailed discussion of the project's objectives.

He also made reference to the fact that a program completion report is currently being prepared by Speech Communications Research Laboratory. This report should be available within a year.

PRECEDING PAGE BLANK NOT FILMED

PRECEDING PAGE BLANK NOT FILMED

PAGE 8 INTENTIONALLY BLANK

(This page intentionally left blank)

D1

[redacted]

N93-72613

176337

# MULTI-SYSTEM APPROACH TO SPEECH UNDERSTANDING

DR. RAJ REDDY

CARNEGIE-MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA

## INTRODUCTION

In 1971, a group of scientists recommended the initiation of a five-year research program towards the demonstration of a large-vocabulary connected speech understanding system (Newell et al., 1971). Instead of setting vague objectives, the group proposed a set of specific performance goals (see Fig. 1.1 of Newell et al., 1971). The system was required to accept connected speech from many speakers based on a 1000 word vocabulary task-oriented grammar, within a constrained task. The system was expected to perform with less than 10% semantic errors, using about 300 million instructions per second of speech (MIPSS)* and to be operational within a five year period. The proposed research was a highly ambitious undertaking, given the almost total lack of experience with connected speech systems at that time.

The Harpy and Hearsay-II systems developed at Carnegie-Mellon University had the best overall performance at the end of the five year period. Figure 1 illustrates the performance of the Harpy system relative to the original specifications. It not only satisfies the original goals, but exceeds some of the stated objectives. It recognizes speech from male and female speakers using a 1011-word-vocabulary document retrieval task. Semantic error is 5% and response is an order of magnitude faster than expected. The Hearsay-II system achieves similar accuracy and runs about 2 to 20 times slower than Harpy.

Of the many factors that led to the final successful demonstration of these systems, perhaps the most important was the systems development methodology that evolved. Faced with prospects of developing systems with large number of unknowns, we opted to develop several intermediate "throw-away" systems rather than work towards a single carefully designed ultimate system. Many dimensions of these intermediate systems were deliberately finessed or ignored so as to gain deeper understanding of some aspect of the overall system. The purpose of this paper is to illustrate the incremental understanding of the solution space provided by the various intermediate systems developed at CMU.

*The actual specifications stated "a few times real-time" on a 100 MIPS (Million instructions per second) machine.

| GOAL (Nov. 1971) | HARPY (Nov. 1976) |
|---|---|
| Accept connected speech | Yes |
| from many | 5 (3 male, 2 female) |
| cooperative speakers | yes |
| in a quiet room | computer terminal room |
| using a good microphone | close-talking microphone |
| with slight tuning/speaker | 20-30 sentences/talker |
| accepting 1000 words | 1011 word vocabulary |
| using an artificial syntax | avg. branching factor = 33 |
| in a constraining task | document retrieval |
| yielding 10% semantic error | 5% |
| requiring approx. 300 MIPSS* | requiring 28 MIPSS |
|  | using 256k of 36 bit words |
|  | costing $5 per sentence processed |

*The actual specifications stated "a few times real-time" on a 100 MIPS (Million instructions per second) machine.

Figure 1. Harpy Performance Compared to Desired Goals

Figure 2 illustrates the large number of design decisions which confront a speech understanding system designer*. For each of these 10 to 15 design decisions, we have 3 to 10 feasible alternative choices. Thus the solution space for speech systems seems to contain $10^6$ to $10^8$ possible system designs. Given the interactions between design choices, it is not possible to evaluate each design choice in isolation outside the framework of the total system.

## SYSTEMS

Figure 3 shows the genealogy of the speech understanding systems developed at CMU. In this section we will briefly outline the interesting aspects of each of these systems and discuss their contributions towards the development of speech understanding systems technology. More complete descriptions of these systems can be found in the references listed at the end.

### THE HEARSAY-I SYSTEM (Erman, Fennel, Lowerre, Neely, and Reddy)**

Hearsay-I (Reddy, Erman, and Neely 1973; Reddy, Erman, Fennel and Neely 1973), the first speech understanding system developed at Carnegie-Mellon University, was demonstrated in June of 1972. This system was one of the first connected speech understanding systems to use task dependent knowledge to achieve reduction of the search space. Recognition uses a best-first search strategy.

### Model

Hearsay-I was the first system to utilize independent, cooperating knowledge sources and the concept of a global data base, or "blackboard", through which all knowledge sources communicate. Knowledge sources consist of the acoustic-phonetic, syntactic, and semantic modules. Each module operates in the "hypothesize-and-test" mode. Synchronous activation of the modules leads to a best-first search strategy. Several other systems have used this stratety (Forgie 1974). This system was one of the first to use syntactically derived word diagrams and trigrams, as anti-productions (Neely 1973), to predict forward and backward from "islands of reliability". Task dependent knowledge, such as a board position in the chess task, is used by the semantic module (Neely 1973), to reject meaningless partial parses early in the recognition process.

*Further discussion of many of these design choices can be found in Reddy (1976).

**The principle contributors towards the development of each of these systems are listed within parentheses.

- Task characteristics
    - speakers; number, male/female, dialect
    - vocabulary and syntax
    - response desired
- Signal gathering environment
    - room noise level
    - transducer characteristics
- Signal transformations
    - digitization speed and accuracy
    - special-purpose hardware required
    - parametric representation
- Signal-to-symbol transformation
    - segmentation?
    - level transformation occurs
    - label selection technique
    - amount of training required
- Matching and searching
    - relaxation: breadth-first
    - blackboard: best-first, island driven
    - productions: best-first
    - Locus: beam search
- Knowledge source representation
    - networks
    - procedures
    - frames
    - productions
- System organization
    - levels of representation
    - signal processor/multi-processor

Figure 2. Design Choices for Speech Understanding Systems

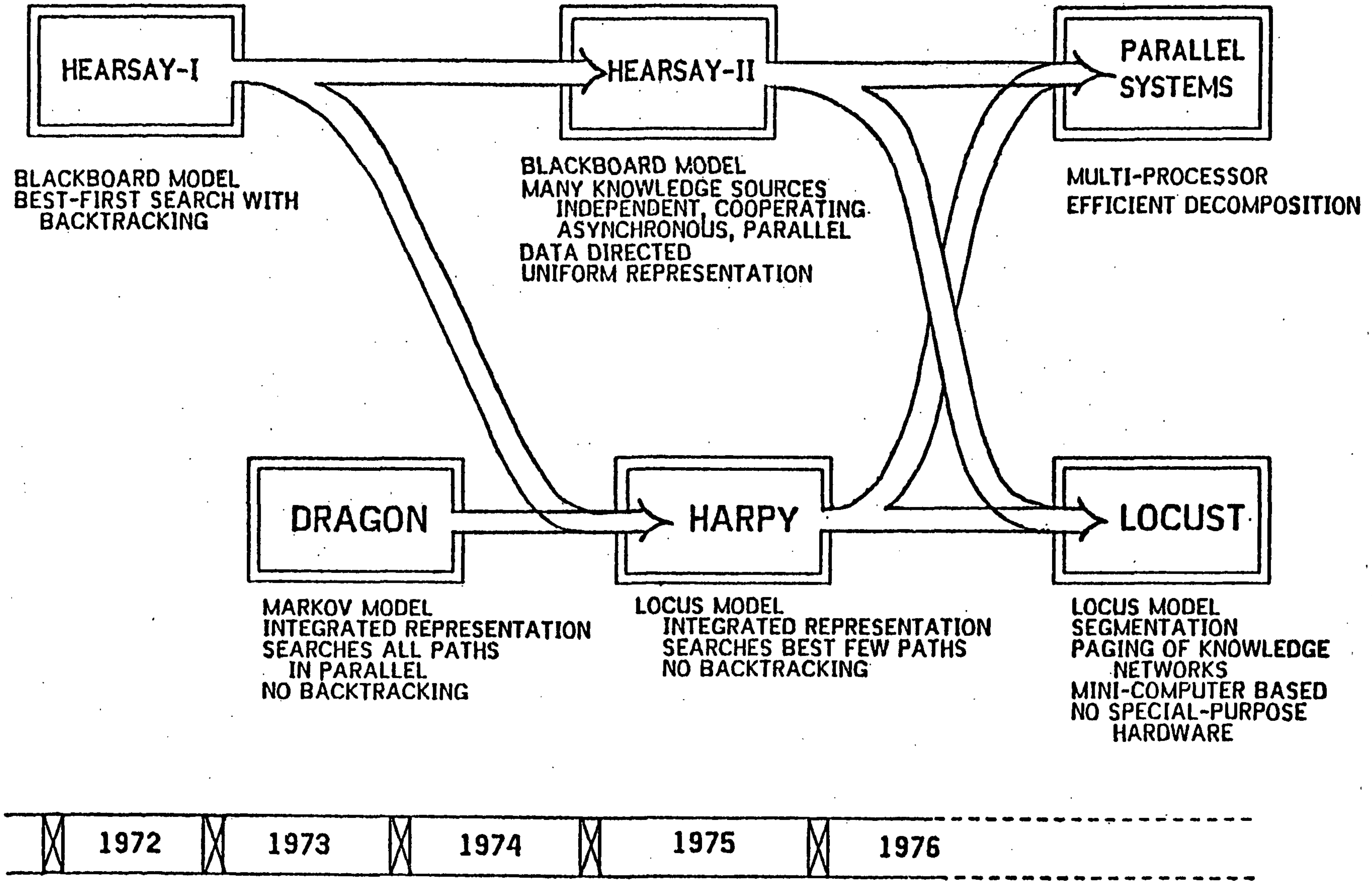


Figure 3. CMU Speech Understanding Systems Genealogy

The acoustic-phonetic module uses amplitude and zero-crossing parameters to obtain a multilevel segmentation into syllable-size and phoneme-size units (Erman, 1974).

### Performance

Over a wide range of tasks, the average sentence error rate was 69% with a word error rate of 45%. Speed varied between 3 and 15 MIPSS over 162 utterances containing 578 words. Hearsay-I yields much higher accuracies on tasks with which it is carefully trained. For the chess task, for instance, average sentence and word error rates were 21 and 7 percent, respectively, with an average speed of 2 MIPSS.

### Discussion

Hearsay-I, as a successful connected-speech understanding system, served to clarify the nature and necessary interaction of several sources of knowledge. Its flexibility provided a means for testing and evaluating competing theories, allowing the better theories to be chosen as a basis for later systems. In retrospect, we believe this system organization would have been adequate for the ARPA specifications given present acoustic-phonetic knowledge.

## THE DRAGON SYSTEM (Baker)

Baker formulated the recognition process as a dynamic programming problem. The Dragon recognition system (Baker, 1975), based on this model was first demonstrated in April of 1974. The system was motivated by a desire to use a general abstract model to represent knowledge sources. The model, that of a probabilistic function of a Markov process, is flexible and leads to features which allow it to function despite high error rates. Recognition accuracy was greater with Dragon than with Hearsay-I, but the system ran significantly slower.

### Model

Dragon was the first system to demonstrate the use of a Markov model and dynamic programming in a connected speech understanding system. It included several interesting features, such as delayed decisions and integrated representation, and is based on a general theoretical framework. The general framework allows acoustic-phonetic, syntactic, and semantic knowledge to be embodied in a finite-state network. Each path through this precomplied network represents an allowed pronunciation of a syntactically acceptable sentence. Recognition proceeds left-to-right through the network, searching all possible paths in parallel to determine the globally optimal path (i.e., the path which best matches the spoken utterance). Acoustic inputs are peak-to-peak amplitudes and zero-crossings from overlapping, one-third octave filters, sampled every centisecond.

### Performance

Recognition accuracy was greater with Dragon than that obtained with Hearsay-I, but at a cost of speed, Dragon being approximately 5 to 10 times slower. Over a wide variety of tasks, the average sentence error rate was 51%. Speed ranged from 14 to 50 MIPSS. The computation is essentially linear with the number of states in the Markov network. Performance was later improved by Lowerre.

### Discussion

Dragon, with more accurate performance than Hearsay-I, served to stimulate further research into factors that led to its improved performance. Many of the ideas motivating its design were important in the development of subsequent connected-speech understanding systems. Although later systems do not use the Markov Model and do not guarantee finding the globally optimal path, the concepts of integrated representation of knowledge sources and delayed decisions proved to be very valuable.

## THE HARPY SYSTEM (Lowerre and Reddy)

The Harpy System (Lowerre 1976) was the first connected speech system to satisfy the original specifications given in the Newell report and was first demonstrated in September of 1976. System design was motivated by an investigation of the important design choices contributing to the success of the Dragon and Hearsay-I systems. The result was a combination of the "best" features of these two systems with additional heuristics to give high speed and accuracy.

### Model

The Harpy system uses the locus model of search. The locus model of search, a very successful search technique in speech understanding research, is a graph-searching technique in which all except a beam of near-miss alternatives around the best path are pruned from the search tree at each segmental decision point, thus containing the exponential growth without requiring backtracking. This technique was instrumental in making Harpy the most successful connected speech understanding system to date. Harpy represents syntactic, lexical, and juncture knowledge in a unified network as in Dragon, but without the a-priori transition probabilities. Phonetic classification is accomplished by a set of speaker-dependent acoustic-phonetic templates based on LPC parameters which represent the acoustic realizations of the phones in the lexical portion of the network.

## Performance

The system was tested on several different tasks with different vocabularies and branching factors. On the 1011-word task the system word error rate was 3% and the semantic error rate was 5% (see fig. 1). The system was also tested with connected digits recognition attaining a 2% word error rate. Using speaker-independent templates, error rate increases to 7% over 20 speakers including 10 new speakers. Using telephone input increases the error rate from 7% to 11% depending on the noise characteristics of the telephone system.

## Discussion

Backtracking and redundant computation have always been problematic in AI systems. The Harpy system eliminates these in an elegant way, using the beam search technique. By compiling knowledge ahead of time, Harpy achieves a level of efficiency that is unattainable by systems that dynamically interpret their knowledge. This permits Harpy to consider many more alternatives and deal with error and uncertainty in a graceful manner.

## THE HEARSAY-II SYSTEM (Erman, Hayes-Roth, Lesser and Reddy)

Hearsay-II has been the major research effort of the CMU speech group over the last three years. During this period, solutions were devised to many difficult conceptual problems that arose during the implementation of Hearsay-I and other earlier efforts. The result represents not only an interesting system design for speech understanding but also an experiment in the area of knowledge-based systems architecture. Attempts are being made by other AI groups to use this type of architecture in image processing and other knowledge-intensive systems.

Hearsay-II is similar to Hearsay-I in that it is based on the hypothesize-and-test-paradigm, using cooperating independent knowledge sources communicating through a global data structure (blackboard). It differs in the sense that many of the limitations and shortcomings of Hearsay-I are resolved in Hearsay-II.

Hearsay-II differs from the Harpy system in that it views knowledge sources as different and independent and thus cannot always be integrated into a single representation. Further, it has as a design goal the ability to recognize, understand, and respond even in situations where sentences cannot be guaranteed to agree with some predefined, restricted language model as is the case with the Harpy system.

Model

The main features of the Hearsay-II system structure are: 1) the representation of knowledge as self-activating, asynchronous, parallel processes, 2) the representation of the partial analysis in a generalized three-dimensional network; the dimensions being level of representation (e.g., parametric, segmental, syllabic, lexical, syntactic), time, and alternatives, with contextual and structural support connections explicitly specified, 3) a modular structure for incorporating new knowledge into the system at any level, and 4) a system structure suitable for execution on a parallel processing system.

Performance

The present system has been tested using about 100 utterances of the training data for the 1011-word vocabulary task. For a grammar with simple syntax (the same one used by Harpy), the sentence error rate is about 16% (semantic error 16%). For a grammar with more complex syntax the sentence error rate is about 42% (semantic error 26%). The system runs about 2 to 20 times slower than Harpy.

Discussion

Hearsay-II represents an important and continuing development in the pursuit of large-vocabulary speech understanding systems. The system is designed to respond in a semantically correct way even when the information is fuzzy and only partial recognition is achieved. Independent knowledge sources are easily written and added to Hearsay-II; knowledge sources may also be removed in order to test their effectiveness. The Hearsay-II system architecture offers great potential for exploiting parallelism to decrease recognition times and is capable of application to other knowledge-intensive AI problems dealing with errorful domains. Many more years of intensive research would be necessary in order to evaluate the full potential of this system.

THE LOCUST SYSTEM (Bisiani, Greer, Lowerre, and Reddy)

Present knowledge representation and search used in Harpy tend to require much memory and are not easily extendable to very large languages (vocabularies of over 10,000 words and more complex syntax). But we do not view this as an insurmountable limitation. Modified knowledge representation designed for use with secondary memories and specialized paging should overcome this difficulty. In addition, it appears larger-vocabulary speech understanding systems can be implemented on mini-computers without significant degradation in performance. Locust is designed to demonstrate the feasibility of these ideas.

### Model

The model is essentially the same as the Harpy system except, given the limitations of storage capacity of main memory, the knowledge representation has to be reorganized significantly. The network is assumed to be larger than main memory, stored on secondary memory, and retrieved using a specialized paging mechanism. The choice of the file structure representation and clustering of the states into pages of uniform size are the main technical problems associated with the development of this system.

### Discussion

A paging system for the 1011 word vocabulary is currently operational on a PDP-11/40E and has speed and accuracy performance comparable to Harpy on a PDP-10 (KA10). Simulation of various paging models is currently in progress. As memories with decreased access times become available, this class of systems is expected to perform as accurately and nearly as fast as systems requiring no secondary memory.

## PARALLEL SYSTEMS (Feiler, Fennell, Lesser, McCracken, and Oleinick)

Response time for the present systems is usually greater than real-time, with indications that larger vocabularies and more complex syntax will require more time for search. One method of achieving greater speed is to use parallel processing. Several systems designed and developed at CMU exploit multi-processor hardware such as Cmmp and Cm*.

### Models

Several systems are currently under development as part of multi-processor research projects which attempt to explore potential parallelism of Hearsay and Harpy-like systems. Fennell and Lesser (1977) studied the expected performance of parallel Hearsay systems and issues of algorithm decomposition. McCracken (1977) is studying a production system implementation of the Hearsay model. Oleinick (1977) and Feiler (1977) are studying parallel decompositions of the Harpy algorithm. Several of these studies are not yet complete, but preliminary performance results are very encouraging. Oleinick has demonstrated a version of Harpy that runs faster than real-time on Cmmp for several tasks.

### Discussion

The main contribution of these system studies (when completed) will be to show the degree of parallelism which can reasonably be expected in complex speech understanding tasks. Attempts to produce reliable and cost-effective speech understanding systems would require extensive studies in this direction.

# DISCUSSION

In the previous section we have briefly outlined the structure and contributions of various speech systems developed at CMU. In retrospect, it is clear that the slow rate of progress in this field is directly attributable to the large combinatorial space of design decisions involved. Thus, one might reasonably ask whether the human research strategy in solving this and other similar problems can benefit from search reduction heuristics that are commonly used in AI programs. Indeed, as we look around, it is not uncommon to find research paradigms analogous to depth-first exploration, breadth-first with shallow cut-off, backtracking, "jumping-to-conclusions", thrashing, and so on.

Our own research has been dominated by two such paradigms. First is a variant of best-first search: find the weakest link (and thus the potential for most improvement) in the system and attempt to improve it. Second is a variant of the beam search: when several alternative approaches look promising, we use limited parallel search with feed-foward. The systems shown in Figure 3 are examples of this type of system iteration and multi-systems approach.

Many system design decisions require an operational total systems framework to conduct experiments. However, it is not necessary to have a single system that permits all possible variations of system designs. Given enough working components, with well-designed interfaces, one can construct new system variants without excessive effort.

The success of the speech understanding research effort is all the more interesting because it is one of the few examples in AI research of a five year prediction that was in fact realized on time and within budget. It is also one of the few examples in AI where adding additional knowledge can be shown to lead to system speed-up as well as improved accuracy.

We note in conclusion that speech understanding research, in spite of the many superficial differences, raises many of the same issues that are central to other areas of AI. Faced with the problem of reasoning in the presence of error and uncertainty, we generate and search alternatives which have associated with them a likelihood value representing the degree of uncertainty. Faced with the problem of finding the most plausible symbolic description of the utterance in a large combinatorial space, we use techniques similar to those used in least-cost graph searching methods in problem solving. Given the problems of acquisition and representation of knowledge, and control of search, techniques used in speech are similar to most other knowledge intensive systems. The main difference is that given human performance the criteria for success, in terms of accuracy and response time, far exceed the performance requirements of other AI tasks except perhaps vision.

## ACKNOWLEDGMENT

I would like to thank Gary Goodman and Lee Erman for their help and comments in the preparation of this paper.

## REFERENCES

J.K. Baker (1975). "Stochastic Modeling as a Means of Automatic Speech Recognition", Ph.D. Dissertation, Dept. of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

J.K. Baker (1975). "The Dragon System - An Overview", IEEE Trans. Acoustic., Speech, and Signal Processing, Vol ASSP-23, pp. 24-29, Feb. 1975.

J.K, Baker (1975). "Stochastic Modeling for Automatic Speech Understanding", in Speech Recognition, D.R. Reddy, (Ed.), Academic Press, New York, 1975.

Computer Science Speech Group (1976). "Working Papers in Speech Recognition IV - The Hearsay-II System", Tech. Report, Dept. of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

L.D. Erman (1974). "An Environment and System for Machine Understanding of Connected Speech", Ph.D. Dissertation, Computer Science Dept., Stanford University, Technical Report, Computer Science Dept., Carnegie-Mellon University, Pittsburgh, PA.

R.D. Fennell and V.R. Lesser (1977). "Parallelism in AI Problem Solving: A Case Study of Hearsay-II", IEEE Trans. on Computers, C-26, pp. 98-111, Feb. 1977.

J.W. Forgie (1974). "An Overview of the Lincoln Laboratory Speech Recognition System", J. Acoust. Soc. Amer., Vol. 56, S27(A).

V.R. Lesser, R.D. Fennell, L.D. Erman, and D.R. Reddy (1975). "Organization of the Hearsay-II Speech Understanding System", IEEE Trans. ASSP-23, No. 1, pp. 11-23.

B.T. Lowerre (1976). "The HARPY Speech Recognition System", Ph.D. Dissertation, Dept. of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

D. McCracken (1977). "A Parallel Production System for Speech Understanding", Ph.D. Thesis (in preparation), Comp. Sci. Dept., Carnegie-Mellon University, Pittsburgh, PA.

R.B. Neely (1973). "On the Use of Syntax and Semantics in a Speech Understanding System", Ph.D. Dissertation, Stanford University, Technical Report, Computer Science Dept., Carnegie-Mellon University, Pittsburgh, PA.

A. Newell, J. Barnett, J. Forgie, C. Green, D. Klatt, J.C.R. Licklider, J. Munson, R. Reddy, and W. Woods, Speech Understanding Systems: Final Report of a Study Group. North-Holland, 1973. Originally appeared in 1971.

D.R. Reddy, L.D. Erman and R.B. Neely (June 1973). "A Model and a System for Machine Recognition of Speech", IEEE Trans. Audio and Electroacoustics Vol. AU-21, (3), pp. 229-238.

D.R. Reddy, L.D. Erman, R.D. Fennell and R.B. Neely (1973). "The HEARSAY Speech Understanding System: An Example of the Recognition Process", Proc. 3rd Int. Joint Conf. on Artificial Intelligence, Stanford, CA., pp. 185-193.

D.R. Reddy (1976). "Speech Recognition by Machine: A Review", Proc. of the IEEE, Vol. 64, pp. 501-531, May 1976.

(This page intentionally left blank)

[redacted]

# CONTRIBUTIONS OF SPEECH SCIENCE TO THE TECHNOLOGY OF MAN-MACHINE VOICE INTERACTIONS

WAYNE A. LEA

SPEECH COMMUNICATIONS RESEARCH LABORATORY, INC.
SANTA BARBARA, CALIFORNIA

## ABSTRACT

Previous interdisciplinary research at Speech Communications Research Laboratory has dealt with a variety of topics in linguistics, speech physiology, perception, and acoustics, plus the interactions among those disciplines. Linear prediction and prosodic correlates of linguistic structures are two examples of research topics that have led to many practical contributions in such application areas as speech recognition. Work in speech recognition has included techniques for vowel identification and normalization, locating syllables, detecting stresses and phrase boundaries, accurately transcribing speech, developing and applying phonological rules, and participating in various aspects of the ARPA SUR project.

Currently a review of the ARPA SUR project and a survey of the speech understanding field are being conducted, with recommendations forthcoming regarding future needs. Several presentations and publications, including a forthcoming book, will report such work. Future plans include prosodics research, phonological rules for speech understanding systems, and continued interdisciplinary phonetics research. One outstanding conclusion from the current review and survey is a renewed call for improved acoustic phonetic analysis capabilities in speech recognizers.

Submitted for publication in the Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command and Control Systems Application, NASA, Ames Research Center, Moffett Field, California.

## 1. Introduction

Speech Communications Research Laboratory (SCRL) is a non-profit research laboratory that was established on the premise that the experimental and theoretical study of spoken language is not simply an adjunct to some other discipline such as electrical engineering or linguistics, but rather it is a distinct and major field of investigation.

[redacted]

It is difficult and, we believe, undesirable to separate our work in speech recognition from the many other disciplines and speech communication problems with which SCRL has worked. This paper consequently begins with a review of the wide range of speech communications projects SCRL has undertaken (section 2). Rather than simply list the many projects, I have organized them within a framework which graphically illustrates the interactions between speech acoustics, physiology, and linguistics. I also offer two examples, concerned with linear predictive analysis and prosodic correlates of linguistic structures, that illustrate how techniques that are directly applicable to speech understanding systems actually originate from interdisciplinary experimental and theoretical research, and then can be turned around to offer evidence for significant changes and new efforts in theory and experimentation.

In section 3, I complete the review of previous SCRL work by briefly describing specific studies in speech recognition that have been conducted at SCRL. These include a number of modest efforts in technology development, and a large project of participation in the Speech Understanding Research ("ARPA SUR") Project sponsored in 1971-1976 by the Advanced Research Projects Agency of the Department of Defense.

Turning from past (Pre-FY '78) work to present and future (Post-FY ' 77) efforts, in section 4 I describe a current contract Dr. June E. Shoup and I are directing, to review the entire ARPA SUR project, to survey all the current technology in speech understanding, and to offer recommendations for further work. This Tri-Services sponsored contract is directly in line with the purposes of this workshop, and should be of widespread interest. We are planning to publish several papers, present several conference talks, and edit two books about speech recognition work throughout the world, and so these outcomes from our project are described in section 5. It is also our hope that from this workshop, from our review, and from related cooperative efforts can come a cataloging of available speech recognition tools, speech databases, and general laboratory facilities for speech analysis, transcription of speech, and collecting statistics about speech regularities. This I discuss briefly in section 6.

Finally, in section 7, I outline our plans for future work on speech understanding.

## 2. The Practical Utility Of Interdisciplinary Research

An understanding of the mechanisms and structures which underly speech is essential to effective man-machine voice communication. We need to call upon the expertise of linguists, phoneticians, engineers.

psychologists, physiologists, speech clinicians, computer scientists, and many other disciplines. For example, it was the psychologists that in recent years clearly demonstrated that no single modality of human communication is as effective in practical problem solving as speech, and that speech is the essential ingredient of the most effective multimodality communication links (Chapanis, 1975).

Engineers and mathematicians gave us the array of valuable speech analysis tools ranging from microphones and electronic filters to Fourier analysis capabilities, fast Fourier transforms, linear predictive analysis, and many other practical devices and algorithms. Computer scientists have given us that fast and versatile tool, the general purpose digital computer, and all its special purpose versions and peripheral devices. More recently, the computer scientists and artificial intelligence advocates have given us practical and effective methods for answering the twenty-year-old call for use of higher-level linguistics knowledge (phonological rules, lexicons, syntax, semantics, and pragmatics) in speech recognition (Denes, 1957; Lindgren, 1965). Decades of work and ideas in acoustic-phonetics, articulatory phonetics, and perception have brought us the phones, phonemes, manner-and-place-of-articulation features, coarticulation constraints, and guidelines about which acoustic changes are truly important (i.e. perceptible), upon which almost all speech recognition and synthesis work is based. Prosodics, as the study of stress, intonation, and the rhythm and timing of speech, had for decades been the concern of comparatively few isolated speech scientists and language teachers, but has recently become one of the prominent subjects in work on speech synthesis and recognition. And so the listing could continue, showing repeated ways in which today's technology builds on yesterday's interdisciplinary science and creative thought. Recently, the ARPA SUR project showed that such a variety of disciplines could work together effectively to develop powerful systems that can successfully understand spoken sentences.

SCRL has, since its founding in 1966, been concerned with the scientific study of the basic linguistic structures of spoken languages, and with the application of this information to problems in electronic communication and speech automation. Gordon Peterson, Founding President and first Director of SCRL, said at the time of SCRL's formation:

> " It is the purpose of the Laboratory to provide a place where scientists and scholars from various disciplines, both technical and humanistic, can work together in mutal respect and enthusiasm on the endless and fascinating problems of speech communication. "

Since that challenging call in 1966, SCRL has been living up to its general goals of discovering basic processes underlying speech communication and sharing the resulting information in the public interest.

While it is recognized that many contributions from basic research do not have widespread impact for many years after the laboratory research is accomplished, it is SCRL's policy to do basic research with specific applications in mind. The result has been that some outstanding ideas and developments at SCRL have had almost immediate direct benefits in practical applications. Perhaps one of the best known examples would be the leading theoretical work of John Markel and his colleagues (Markel, 1972; Markel and Gray, 1973; 1974; 1976) on linear predictive analysis, which has already been applied in systems for speech recognition, speaker authentication or identification, and early detection of laryngeal cancer. Markel is currently applying his techniques to government applications in speaker recognition, within a newly formed applications-oriented company he directs. His linear predictive coding techniques have also been adopted by many other groups working on speech analysis and synthesis throughout the world. If someone had stopped that type of rigorous mathematical work at its early stages only a few years ago, on the mistaken notion that it was irrelevant to immediate practical needs, where would our speech analysis and synthesis capabilities be today? We might still be struggling to extract the really important spectral cues (formants, fundamental frequency, glottal waveforms, vocal tract area functions, et.) from the complicated, noisy speech spectra that for twenty or more years had defied reliable automatic analysis.

Linear predictive analysis is a good model for illustrating the interdisciplinary origins and applicabilities of speech research. The mathematical models, that are now implementable in practical forms in general purpose (or specialized) computers, have been shown to be appropriate to capture the essence of the accustic modulation of a vocal-cord source that is produced by the variable-cross-section vocal tract. Linear prediction permits detection of vocal tract resonances (formants or transfer-function poles), voice fundamental frequency and waveforms of airflow at the vocal cords, and radiation impedance at the lips. It is known to be appropriate for vowels and oral consonants, and even though our knowledge of articulation and acoustic phonetics suggests its mathematical inapplicability for nasal consonants, practical approximations and perceptual significances tell us that it is possible to learn something about the speech (e.g., approximate nasal resonances and bandwidths) even when the model's mathematical assumptions are not strictly met. Here we see acoustics, articulatory phonetics, perception, linguistic category distinctions, mathematics, computer science, and practical engineering approximations all coming into play. Then we see linear predictive anaylsis used to aid vowel and consonant identification in speech recognition, plus detect talker-specific differences in vocal tracts and voice sources, and even detect laryngeal cancer and other speech pathologies. One recent project at SCRL used the residual energy function from a linear predictive analysis to detect laryngeal (voice) pathologies such as cancer, and to provide "voice profiles" that may be useful in clinical, musical, and legal applications (Davis, 1976)

Another example of interdisciplinary interactions is my own growing interest in prosodic structure. When, in 1966, Gordon Peterson and his colleagues at SCRL first introduced me to the obscure area of phonetics and linguistic studies they called "prosodic structures", I had no idea how prosodics studies would lead to such a variety of scientific questions and practical applications. Following the linguists' arguments that stress patterns are determined by the phrase structures of sentences, and the phoneticians' studies of acousitc prosodic correlates of stress, I hypothesized that one should be able to determine aspects of syntactic structure directly from acoustic prosodic features. This led to the development of a computer program which detected about 90% of major phrase boundaries in connected speech, using only fall-rise valleys in intonation patterns. Another program detected syllabic nuclei from bandlimited energy functions, and used energy, syllabic durations, and fundamental frequency contours to successfully locate about 90% of the stressed syllables. Extensive series of experiments were conducted on the intonation, perceived stress patterns, rhythms, and pauses in various speech texts. Methods were devised for using such prosodic information to aid phonemic analysis, word matching, and parsing in speech understanding systems. In fact, a general prosodically-guided speech understanding system strategy was outlined, and aspects of it were incorporated into the developing system at Sperry Univac (Lea, Medress, and Skinner, 1975).

All this prosodics research which I did while at Sperry Univac is summarized in a recent report (Lea, 1977). It clearly showed the potential for extracting aspects of syntactic structure from acoustic prosodic data, independent of any knowledge of the wording of the sentence. Prosodics also can be used to reduce the set of alternative words that should be hypothesized at each point in an unknown utterance. Hypothesized words should have stresses expected where they are actually found in the acoustic prosodic data (for example, word-finally stressed "abridge" should not be hypothesized or should be given a lower priority for testing where the prosodics clearly suggest an initially-stressed word like "average"). Also, only certain words can be in phrase-initial or phrase-final positions, so if a phrase boundary is reliable detected, one can confine hypothesized words to those that could appear in those patterns.

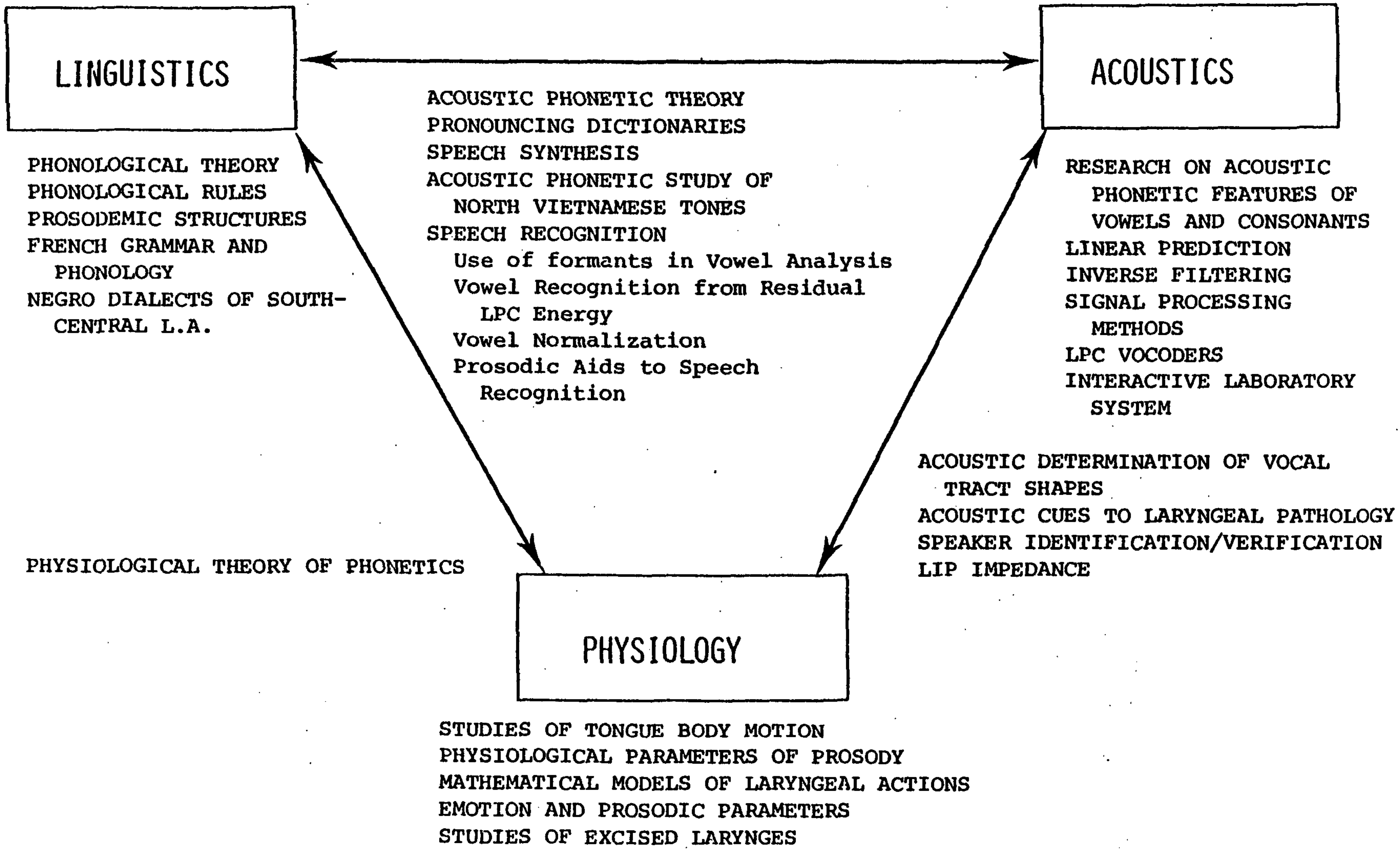
Those prosodic studies, which began from general linguistic theories and acoustic phonetic experiments, thus developed into substantial contributions to practical aspects of computer understanding of spoken sentences. Then, as if to complete the circle, some of the acoustic prosodic features detected in such analyses led to widespread theoretical implications, such as explanations for how tones develop or disappear in the historical change of a language (or family of languages), how consonants interact with tones in tone languages, why stresses tend to be equally spaced (isochronous) in English, which of the linguist's stress rules are evident in acoustic data and listeners perceptions of stress, etc. I also used available automatic phonetic analysis routines to confirm a long-held notion that stressed syllables provide "islands of phonetic reliability" in speech. These studies also raised questions about the physiological origins of higher fundamental frequencies in high (vs. low) vowels, relationships between larynx height and fundamental frequency, the physiological origin of gradually falling intonation, etc.

We thus have two quite different examples of practical benefits coming from some interdisciplinary research. A detailed discussion of other SCRL interdisciplinary work is impossible here, but we can list many of the other topics that have been studied, and indicate some structure for relating all these studies to each other and to our main topic of speech recognition.

Gordon Peterson characterized the interrelationships between acoustics, physiology, and linguistics by the basic triangle shown in Figure 1. I have illustrated on the diagram the various topics of research to which SCRL has contributed during its various government-sponsored and privately funded contracts and grants. This listing of topics was compiled from the list of over 100 journal articles, book chapters, and reports, plus 14 books and monographs, that SCRL researchers have published. The work ranges from abstract linguistic studies like grammar, phonology, dialects, and abstract prosodic ("prosodemic") structures, to extensive studies of acoustic features of vowels and consonants, and a variety of signal processing techniques and applications. Physiology, as something of a "way station" between linguistics and acoustics, has been the subject of several medical studies and some mathematical modelling at SCRL.

Outstanding among the published works from SCRL are Peterson and Shoup's "Physiological and Acoustic Theories of Phonetics" (1966). These links between linguistics and either acoustics or physiology are shown by the top and left arrows in Figure 1. Also linking linguistics and acoustics are developments of dictionaries specifying the actual ways words are pronounced in various forms of communication (read speech, formal talks, conversation, etc.). Speech synthesis is an "encoding" effort, which allows going from specified linguistic messages to automatically composed acoustic forms that are acceptable and intelligible to listeners. Speech recognition, the primary topic of the remainder of this paper, is the opposite process of automatically determining linguistic messages from acoustic data.

Many researchers have noted the difficulty of relating accoustic data to underlying abstract linguistic messages, and acknowledged the importance to be attached to the fact that speech is produced by very specific physical mechanisms that are more readily accessible than neural



PHONOLOGICAL THEORY
PHONOLOGICAL RULES
PROSODEMIC STRUCTURES
FRENCH GRAMMAR AND PHONOLOGY
NEGRO DIALECTS OF SOUTH-CENTRAL L.A.

ACOUSTIC PHONETIC THEORY
PRONOUNCING DICTIONARIES
SPEECH SYNTHESIS
ACOUSTIC PHONETIC STUDY OF NORTH VIETNAMESE TONES
SPEECH RECOGNITION
Use of formants in Vowel Analysis
Vowel Recognition from Residual LPC Energy
Vowel Normalization
Prosodic Aids to Speech Recognition

RESEARCH ON ACOUSTIC PHONETIC FEATURES OF VOWELS AND CONSONANTS
LINEAR PREDICTION
INVERSE FILTERING
SIGNAL PROCESSING METHODS
LPC VOCODERS
INTERACTIVE LABORATORY SYSTEM

PHYSIOLOGICAL THEORY OF PHONETICS

ACOUSTIC DETERMINATION OF VOCAL TRACT SHAPES
ACOUSTIC CUES TO LARYNGEAL PATHOLOGY
SPEAKER IDENTIFICATION/VERIFICATION
LIP IMPEDANCE

STUDIES OF TONGUE BODY MOTION
PHYSIOLOGICAL PARAMETERS OF PROSODY
MATHEMATICAL MODELS OF LARYNGEAL ACTIONS
EMOTION AND PROSODIC PARAMETERS
STUDIES OF EXCISED LARYNGES

Figure 1. Speech Research Topics Studied in Previous Work at SCRL

commands of linguistic import. Consequently, physiology has played a major role in speech anlysis studies. In particular, it is frequently noted in speech recognition studies that manner of articulation (that is, whether a particular segment of speech is a vowel, a stop consonant, a fricative, a nasal consonant, or what) is more easily and reliably determined than place of articulation (such as, at the teeth, at the alveolar ridge, near the velum, etc.). Similarly, the physiological differences between male and female talkers is a notable reason for significant acoustic differences in their spoken vowels and consonants. The automatic recognition of voices is a way of linking acoustics to physiology. Two of the most impressive recent developments in speech science are concerned with (a) determining the vocal tract shape and (b) detecting laryngeal pathologies (such as cancer of the larynx), both directly from acoustic features. Major work in these areas was done at SCRL (Wakita, 1973, Davis, 1977).

While all this work impinges upon methods for speech recognition, there are some specific recognition projects that will be given special attention in the next subsection, to complete this review of previous (Pre-FY '78) work at SCRL.

### 3. Speech Recognition Studies at SCRL

Speech recognition research has been an important part of the projects and interests of the staff of SCRL since even before the founding of SCRL in 1966. In the late 1950's and early 1960's, while he was still with Bell Telephone Laboratories and the University of Michigan, Gordon Peterson outlined general models of automatic speech recognition and called for the use of linguistic structures, prosodics, and articulatory-based models to augment incoming acoustic information. Peterson was a leader in acoustic phonetic research and the author of works that are still among the most widely quoted in the field (e.g., Peterson and Barney, 1952). At the Univeristy of Michigan in 1963, he and Dr. June E. Shoup, the present Director of SCRL, conducted an epic-making course in Automatic Speech Recognition involving outstanding leaders in various related fields.

SCRL staff members have written several foundational papers concerning basic methods in speech recognition (Shoup; 1968, Broad, 1972 a,b; Broad and Shoup, 1975; Broad, 1976). In a frequently referenced paper, Broad (1972 a) described how to use formants in automatic speech recognition. Pilot experiments were also done on using residual energy of a linear prediction analysis to identify vowels. A method was developed for speech segmentation and normalization of spectral features based on the acoustically-derived vocal tract area functions (Kasuya and Wikita, 1976) and vocal tract length (Wakita, 1977). Automatic detection of syllabic nuclei was also studied at SCRL (Wakita and Kasuya, 1977).

The largest long-term effort in speech recognition at SCRL was undertaken within the ARPA SUR project. As a Support Contractor, SCRL developed new analysis tools and provided a variety of services for speech understanding system builders, such as:

- Doing a well-controlled phonemic analysis of a common database of "31 ARPA test sentences";
- Compiling lists of phonological rules;
- Developing methods for generating small dictionaries from lists of words related to a speech understanding task;
- Studying the feasibility of a common task for direct comparison of alternative speech understanding systems;
- Relating the literature on the location of syllable boundaries to the formal statements of phonological rules;
- Transcribing large speech databases orthographically, phonemically, and phonetically;
- Participating in planning meetings and workshops in acoustic parameterization, phonemic segmentation and labeling, and phonology.

SCRL cooperated with SDC, CMU, and BBN in their efforts to compile speech databases, develop and test segmentation and labeling schemes, and implement baseform dictionaries and phonological rules. My own work on prosodic aids to speech recognition, while initially done at Sperry Univac, may also now be considered part of the SCRL background in automatic speech recognition.

In summary of the SCRL work before FY '78, we have seen that general speech sciences work in linguistics, physiology, and acoustics, and the ties between those disciplines, have provided a general interdisciplinary background for a variety of specific studies in speech recognition. SCRL's specific ASR studies have ranged from detailed analysis and identificiation of vowels (using formants, residual LPC energy functions, and/or vocal tract area functions) to general theories of automatic speech recognition and rules for phonological anlysis. The pronouncing dictionary at SCRL is very large (300,000 entries), and orthographic, phonemic, and phonetic transcription methods are highly developed, and have been extensively used, at SCRL and by speech understanding system builders.

### 4. Tri-Services Contract to Review ARPA SUR and Survey the Current Technology

On July 20, 1977 SCRL was awarded a contract, sponsored by the Tri-Services and the Advanced Research Projects Agency, to review the five-year, $15-million ARPA SUR project and to survey the current technology in speech understanding. One task is to review and evaluate the performance of the speech understanding systems developed by Bolt Beranek and Newman (BBN), by the speech group at Carnegie Mellon University (CMU), and by the Systems Development Corporation (SDC, in cooperation with the Stanford Research Institute). We have read the various reports prepared by these groups, and have visited their laboratories to discuss the structures of their systems, the final performance results, their assessments of various aspects of their work, and their judgments about what work should now be done on speech understanding systems. We have concentrated on the techniques they consider to have been particularly successful, and have discussed with them the weakest points of their systems, and what further work is consequently needed. We have tried to understand why some systems have succeeded more than others, and have discussed what work these groups would want to do if given either one year or five years of further opportunity to extend their work. This provided us with a catalog of suggestions about work that deserves immediate attention, and work that should be included in the next major advance in speech understanding technology.

The significance of such a study can hardly be overemphasized. When ARPA initiated the ARPA SUR project over five years ago, the objective was to obtain a breakthrough in the ability of computers to understand spoken sentences. During two decades of prior research there had been repeated calls for overcoming the major hurdle separating moderately successful isolated-word-recognition systems from the unattained ideal of more natural uninterrupted voice communication with computers. Review articles had repeatedly called for the full use of language structures such as acoustic phonetics, coarticulation regularities, phonological rules, prosodic structures, syntax, and semantics (Lindgren, 1965; 1965; Hill, 1971; Lea, 1972; Broad, 1972 b). The ARPA project was the first large-scale effort to provide such a technology for understanding spoken sentences.

The original study report which formed the blueprint for the ARPA SUR project (Newell, et al., 1971) noted that successful speech understanding by computer depends on integrating various types of knowledge (e.g., acoustics, phonetics, syntax, etc.) and applying this multilevel information to the interpretation of utterances within a specific task domain. We are examining how ARPA SUR participants characterized these kinds of knowledge and organized these components into speech understanding systems, and are attempting to evaluate the various components. The original ARPA SUR study group outlined goals that were very ambitious, given the fledgling state of continuous speech recognition and the defensive posture the field had following Pierce's (1969) pessimistic evaluation of speech recognition work (cf. Lea, 1970). Yet, the specific goals of the project are considered to have been substantially met by the HARPY speech understanding system that was demonstrated at Carnegie-Mellon University (CMU) on September 8, 1976. Other systems developed at BBN and SDC also attained some success in sentence understanding, though more ambitious goals of handling a sizeable subset of spoken English and conducting longer-range research appear to have prevented those systems from being tested, refined, and constrained adequately to attain the high (90%) semantic accuracy set down in the original goals. Still, many ideas and implementation techniques have been considered and tested in these systems that should be clearly understood, evaluated, and applied as appropriate in the development of future systems.

In addition to the CMU, BBN, and SDC systems, preliminary systems were developed at Lincoln Laboratory of MIT and Stanford Research Institute, and tested with some success in 1974. Also, supporting speech research efforts were conducted at Haskins Laboratories, Sperry Univac, and the University of California at Berkeley (transferred from the University of Michigan during the project), as well as at SCRL. We are also reviewing the scientific and technological advancements resulting from such work.

A five-year, $15-million, multiple-contractor program the size of the ARPA SUR project certainly deserves careful review and evaluation. Our responsibility as we see it is to evaluate the project with tomorrow in mind, not yesterday, so that we propose to address such questions as the following:

- What were the specific scientific and technological accomplishments in the SUR project?
- How has the state of the art in speech understanding advanced from 1971 to now?
- What problems in speech analysis became apparent from the efforts to provide systems that met the original specifications?
- What type of components produced the best results? The worst results? What are the sources of errors? In particular, what are the most common reasons for a system's being sidetracked into exploring wrong hypotheses about sentence structures?

Our review will hopefully provide an accurate picture of how the ARPA SUR project produced progressive steps in the technology of speech understanding systems. To complete a picture of the state of the art in 1977, we are attempting to relate the performance and techniques of the ARPA SUR systems to other work in the field. As soon as our ARPA SUR review is complete, we will study work at IBM, Sperry Univac, Bell Laboratories, ITT, Texas Instruments, Threshold Technology, and many other groups throughout the world. We hope to determine the adequacies and inadequacies of current capabilities and to help establish what is left to do to produce useful systems for a spectrum of applications. Some of the questions being addressed are:

- Where does the rest of the speech understanding field stand and how do the accomplishments of the ARPA/SUR program fit in with other work?
- What remains to be done to attain useful forms of speech understanding systems for DOD applications?
- How extendable are the current systems? Can they be made to operate with a natural ("habitable") subset of English? What is still needed to provide a spectrum of systems for handling various applications?

There are several dimensions of task difficulty in the speech understanding framework that need to be explored further. What happens to the performance of the alternative systems for speech understanding when:

- The language gets more complex and flexible
- The number of expected talkers increases
- Dialects and speech styles change
- The microphone or communication channel includes noise, bandwidth limitations, distortions, etc.
- The system cannot be as extensively trained (or not trained at all) for each talker
- The practical needs of real time operation on moderate-sized available computers are taken into full consideration
- Real task domains such as applications in the military services are tackled

- Very high accuracy in semantic understanding is demanded.

It is, of course, very difficult to assess the whole technology of speech understanding, and we have not been so presumptuous as to think we can answer all these (and other) questions by ourselves. We have distributed a questionnaire to about 100 researchers and technologists in speech recognition, seeking their opinions about the ARPA SUR project, the current technology, and the future work that is needed.

One of the primary goals of this Tri-Services study is to determine what needs to be done in future work on speech recognition and/ or understanding. In addition to studies of all the documentation from the ARPA SUR project and other current work, and interactions with various workers to define the detailed adequacies and inadequacies of current systems and their components, we would like to work with ARPA and the military services to define what yet needs to be done and where to go from here. We all need the information being given at this workshop about DOD speech recognition applications, gaps in speech recognition capabilities, and possible programs for future development of useful systems.

## 5. Forthcoming Publications and Presentations

A primary outcome from the Tri-Services review and survey will be a series of publications summarizing what we have learned. The following is a list of publications and public presentations that are to appear:

- W. A. Lea and J. E. Shoup, Specific Contributions of the ARPA SUR Project to Speech Science, to be presented at the 94th Meeting of the Acoustical Society of America, Miami, Florida, December 14, 1977. (Abstract in J.A.S.A., vol. 62, Suppl. 1, Fall, 1977).

- W. A. Lea, President of a Special Session on "Speech Recognition: What is Needed Now?", International Phonetic Sciences Congress (IPS-77), Miami, Florida, December 19, 1977.

- J. E. Shoup, "Phonological Aspects of Speech Recognition:, to be presented at the IPS-77 Special Session on "Speech Recognition: What is Needed Now?", Miami, Florida, December 19, 1977.

- W. A. Lea and J. E. Shoup, "Gaps in the Technology of Speech Understanding", to appear in Proc. 1978 IEEE International

Conf. on Acoustics, Speech and Signal Processing, Tulsa, Oklahoma, April 10-12, 1978

- TRENDS IN SPEECH RECOGNITION, a book edited by W. A. Lea, including the following papers by SCRL researchers:

  VOLUME I: (GENERAL ISSUES AND TRENDS)

  Ch. 1. The Value of Speech Recognition Systems (W.A. Lea)
  Ch. 4. Speech Understanding Systems: Past, Present and Future (W.A. Lea)
  Ch. 6. Phonological Aspects of Speech Recognition (J.E. Shoup)
  Ch. 7. Prosodic Aids to Speech Recognition (W.A. Lea)
  Ch. 17. Specific Contributions of ARPA SUR to Speech Science (W.A. Lea and J.E. Shoup)
  Ch. 23. Speech Recognition Work in Asia (H. Wakita and Shuzo Makino)
  Ch. 27. Speech Recognition: What is Needed Now? (W.A. Lea)

- W.A. Lea and J.E. Shoup to conduct a Workshop on Speech Understanding Technology and Its Applications, Washington D.C., Spring, 1978.

- W.A. Lea and J.E. Shoup, Review of the ARPA SUR Project and Survey of the Speech Understanding Field, Final Report on ONR Contract No. N00014-77-C-0570.

- W.A. Lea, "Advances in Speech Recognition", invited paper to appear in Proceedings of the IEEE, Special Issue on Pattern Recognition, May 1979.

- W.A. Lea, "Voice Input to Computers: An Overview", an invited talk to be presented at the National Computer Conference, Anaheim, CA, June 6-8, 1978.

Previous reviews of the ARPA SUR project have concentrated on final system performance and a general description of the systems developed. Our paper for the ASA meeting in Miami is intended to focus attention on the basic speech science results from the project. Only some of these results were actually incorporated into the final systems. Some were excluded in the final rush to complete work on operational but restricted systems, and some scientific contributions by the support contractors were not translated into specific algorithms for use in systems.

Dr. Shoup and I will endeavor to outline those gaps in speech understanding technology that need early attention, based on our survey of the current state of the art. Only some of these gaps can be included in the written version of the IEEE/ICASSP paper, which is due December 19, but more will be included in our oral presentation next April.

Also, in December, I am chairing a session at the IPS-77 Congress, which I have deliberately organized to focus international attention on the current technology and future needs in speech recognition. June E. Shoup is presenting an invited paper at that session on phonological aspects of recognition, which will be based on her review of phonological studies within ARPA SUR and the entire current technology.

The IPS-77 papers from that session, and 20 other papers from the most active groups throughout the USA and the world, will be included in a book which I am editing, and which is scheduled for publication in 1978. There is a section (composed of several papers) covering the ARPA SUR project, several papers on the need for speech recognition, tutorial papers about aspects of speech understanding system design, a series of papers about recent operational systems in the USA, and several survey articles dealing with the work in other countries. Much of our review and survey work is to be included in our chapters in that book. We have also been invited to provide a general review of the field for the Proceedings of the IEEE, a tutorial review for the IEEE Spectrum, and an overview for the National Computer Conference. Our final report will be issued next August, and will include all of our review and survey results, and our recommendations for future work.

## 6. Cataloging Available Services and Tools

Many computer programs have been developed in the course of the ARPA SUR project and other previous work. Extensive sets of sentences have been recorded, digitized, processed for important parameters, segmented and labeled with phonetic or phonemic category symbols. Some sentences have been transcribed by linguists, and in some cases those transcriptions have been time-locked to the speech waveform, so that valuable data for studying the acoustic phonetic, prosodic, and phonological structures of English sentences have been obtained. Also, valuable laboratory facilities have been developed for analyzing speech, playing it back (repeatedly, if desired, as in perception experiments), processing it for parameters, automatically segmentating and labeling, and many other speech-handling tasks. Statistical packages have been developed to keep track of such data, to automatically do analyses of regularities, and to plot such displays as histograms, discrimination thresholds, etc.

All this work should be cataloged and made available to all interested groups (where possible), so that duplication of efforts and costly diversions can be avoided in future studies. We hope to do some of that cataloging as time permits within our contract, and to outline general ways in which organizations like the IEEE Subcommittee on Speech Recognition can make such services and tools available to other researchers and developers of systems.

## 7. Future Plans

Obviously, since we are currently involved in a review and survey that will define what work should be undertaken in future studies, we cannot, and should not, at this time offer detailed plans for future work. We do have some general plans, and ideas for specific work that is in keeping with all that we have learned in our ARPA SUR review, discussions with other researchers, and survey to date. SCRL will continue to be involved in speech understanding studies, since the need for such facilities remains and there are significant gaps still to be filled in the available technology. In particular, we plan to pursue prosodics research and develop an improved and expanding capability in prosodic aids to speech understanding. Prosodics has been one of the knowledge sources that has been most obviously missing from previous systems, not only in our opinion but in the opinions of several other leading groups with whom we have visited (also, cf. Woods, 1974, p. 9; Wolf, 1977, p. 207).

Another major need reiterated by every group we conferred with is improved acoustic phonetic analysis (the so-called "front end" of many systems). SCRL has a long term history in such studies, and will presumably contribute to such work. However, the work in substantially improving acoustic phonetics aspects of recognition is very demanding and will require cooperative efforts by many different research, technology, and applications-oriented groups. It is particularly striking that major improvements in acoustic phonetics capabilities are needed despite decades of excellent work in that field, while ARPA's five year ambitious effort in artificial intelligence and higher level linguistics constraints has achieved such substantial progress that the bottleneck is again back in the difficult problem areas of acoustic segmentation, labeling and preliminaries to word identification.

I also see a definite need for future understanding systems to be tested on a common task (that is, evaluated with the same speech data and task domain) or else evaluated with carefully designed "performance metrics" that make it possible to decide whether 50% correct recognition on a difficult task is better or worse than 95% recognition on a much easier task. This is, for example, relevant in trying to comparatively evaluate the ARPA SUR systems developed at CMU, BBN, and SDC. Very little work has been done on performance metrics and task complexity metrics that can make possible the comparative evaluation of alternative systems (cf. Goodman, 1976; Moore, 1977).

In conclusion, I have listed in Figure 1 the variety of research topics which SCRL has addressed in the past eleven years. I have sought to illustrate, with linear prediction and prosodic aids to speech understanding, some graphic examples of how interdisciplinary speech sciences research can readily lead to a variety of practical tools and provoke further scientific research. SCRL has conducted several studies in speech recognition, including providing transcription capabilities, prosodics research, and phonological analyses for the ARPA SUR project. We are currently engaged in a review of ARPA SUR, a survey of the speech understanding field, and a development of recommendations for future work in the field. We will be reporting our work in a number of publications, and already see several definite areas for further work, including prosodics, task complexity measurement (and performance metrics), and further advances in acoustic phonetic aspects of recognition.

## 8. References

D.J. Broad (1972 a), Basic Directions in Automatic Speech Recognition, Intern. J. Man-Machine Studies, vol. 4, 105-118.

D.J. Broad (1972 b), Formants in Automatic Speech Recognition, Intern. J. Man-Machine Studies, vol. 4, 411-424.

D.J. Broad (1976), Acoustic Discrimination Between (f) and (o) in a Single Speaker, Proc. 1976 IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing, 162-165.

D.J. Broad and J.E. Shoup (1975), Concepts for Acoustic Phonetic Recognition. Speech Recognition, (D.R. Reddy, editor). New York: Academic Press, 243-274.

A. Chapanis (1975), Interactive Human Communication, Scientific American, March 1975, 36-42.

S.B. Davis (1977), Computer Evaluation of Laryngeal Pathology Based on Inverse Filtering of Speech, SCRL Monograph No. 13, SCRL, Santa Barbara, California.

P. Denes (1957), The Design and Operation of the Mechanical Speech Recognizer at University College London, The Journal of British Institution of Radio Engineers, vol. 19, 219-229.

R.G. Goodman (1976), Analysis of Languages for Man-Machine Voice Communication, Ph.D. Dissertation, Computer Sciences Dept., Carnegie-Mellon University, Pittsburgh, Pennsylvania.

D.R. Hill (1971), Man-Machine Interaction Using Speech, *Advances in Computers*, vol. 11, 165-230.

H. Kasuya and H. Wakita (1976), Speech Segmentation and Feature Normalization Based on Area Function, *Proc. 1976 IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, 29-32.

W.A. Lea (1969), Evaluating Speech Recognition Work. *J. Acoust. Soc. of America*, Vol. 47, No. 6, 1612-1614 (L).

W.A. Lea (1972), Intonational Cues to the Constituent Structure and Phonemics of Spoken English, Ph.D. Dissertation School of Electrical Engineering, Purdue University, Lafayette, Indiana.

W.A. Lea (1976), Prosodic Aids to Speech Recognition: IX. Acoustic-Prosodic Patterns in Selected English Phrase Structures, *Univac Report No. PX11963*, Sperry Univac DSD, St. Paul, Minnesota.

W.A. Lea, M.F. Medress, and T.E. Skinner (1975), A Prosodically-Guided Speech Understanding Strategy, *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-23, 30-38.

N. Lindgren (1965), Machine Recognition of Human Language, *IEEE Spectrum*, vol. 2, March and April, 114-136 and 45-59.

J.D. Markel, Automatic Formant and Fundamental Frequency Extraction from a Digital Inverse Filter Formulation, *1972 Inter. Conf. on Speech Communication and Processing*, Boston, Massachusetts, 81-84.

J.D. Markel and A.H. Gray, Jr. (1973), On Autocorrelation Equations with Application to Speech Analysis, *IEEE Trans. on Audio and Electroacoustics*, vol. AU-21, No. 2, 69-79.

J.D. Markel and A.H. Gray, Jr. (1974), A Linear Prediction Vocoder Simulation Based Upon the Autocorrelation Method, *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-22, No. 2, 124-134.

J.D. Markel and A.H. Gray, Jr. (1976), *Linear Prediction on Speech*. Berlin, Heidelberg, New York: Springer-Verlag.

R.K. Moore (1977), Evaluating Speech Recognizers, IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. ASSP-25, No. 2, 178-183.

A. Newell, J. Barnett, J.W. Forgie, C.C. Green, D.H. Klatt, J.C.R. Licklider, J. Munson, D.R. Reddy, W.A. Woods, (1971), Speech Understanding Systems: Final Report of a Study Group, Computer Science Department, Carnegie-Mellon University, Pittsburgh, Pennsylvania.

G.E. Peterson and H. Barney (1952), Control Methods Used in a Study of the Vowels, J. Acoust. Soc. of America, vol. 24, 175-184.

G.E. Peterson and J.E. Shoup (1966), a Physiological Theory of Phonetics, Journal of Speech and Hearing Research, vol. 9, No. 1, 6-67. Also: The Elements of an Acoustic Phonetic Theory, Journal of Speech and Hearing Research, vol. 9, No. 1, 68-99.

J.R. Pierce (1969), Whither Speech Recognition? J. Acoust. Soc. of America, vol. 46, 1049-1051.

J.E. Shoup (1968), Approaches to Automatic Speech Recognition, Naval Reviews, June 1968, 11-17.

H. Wakita (1973), Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms, IEEE Trans. on Audio and Electroacoustics, vol. AU-21, No. 5, 417-427.

H. Wakita (1977), Normalization of Vowels by Vocal-Tract Length and Its Application to Vowel Identification, accepted for publication in IEEE Trans. on Acoustics, Speech, and Signal Processing (in press).

H. Wakita and H. Kasuya (1977), A Study of Vowel Normalization and Identification in Connected Speech, Proc. 1977 Inter. Conf. on Acoustics, Speech, and Signal Processing, 648-651.

J. Wolf (1977), "Speech Recognition"; Invited Papers Presented at the 1974 IEEE Symposium (D.R. Reddy, Ed.). Book Review, in IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. ASSP-25, No. 2, 207.

W.A. Woods (1974), Motivation and Overview of BBN SPEECHLIS: An Experimental Prototype for Speech Understanding Research, Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Pennsylvania, 1-10.

## 9. Acknowledgements

This study is supported by the Tri-Services and the Advanced Research Projects Agency of the Department of Defense and is monitored by the Office of Naval Research under Contract #N00014-77-C-0570.

## BIOGRAPHICAL SKETCH

### Wayne A. Lea

Wayne A. Lea joined the staff of Speech Communications Research Laboratory, Inc. (SCRL) in August, 1977, and is serving as a Research Linguist and Research Engineer. His primary responsibility is as Co-Principal Investigator (with June E. Shoup), on a Tri-Services sponsored contract to review the five-year ARPA sponsored Speech Understanding Research project, to survey the current state of speech understanding technology, and to recommend further work relevant to future DOD applications for speech understanding systems. He also serves as Coordinator of Private Funding at SCRL.

Prior to joining SCRL, Dr. Lea was Co-Principal Investigator at Sperry Univac on a five-year research contract for ARPA, concerned with developing prosodic programs that located syllables, detected intonationally-marked phrase boundaries, and located stressed syllables. He also conducted a series of experiments on: human perception of stress patterns; prosodic correlates or linguistic structures; regularities in English rhythm, pause structures, accent, and intonations; and machine analysis of phrase boundaries, stress patterns, and phonetic structures. He proposed, and, for a time served as Co-Principal Investigator on, a project to spot occurrences of key words in continuous speech. Prior to his work at Sperry Univac, Dr. Lea has conducted prosodic research with the Purdue Research Foundation for two years. He also served four years with NASA at the Electronics research on speech recognition, mathematical linguistics, and the effectiveness of voice as a modality for man-machine interaction. Earlier he has worked on cybernetics, linguistics, and artificial intelligence projects at Montana State College and Massachusetts Institute of Technology.

Dr. Lea earned his Bachelor and Master of Science degrees in Electrical Engineering at Montana State College, then completed an Interdisciplinary Science Master's degree (with linguistics emphasis) and a professional Electrical Engineer degree at Massachusetts Institute of Technology. His doctorate is an interdisciplinary one (Electrical Engineering, English, and Audiology and Speech Sciences) from Prudue University. He has published over 50 reports, journal articles, conference papers, and book chapters, and is currently editing two books.

# RESEARCH ON SPEECH UNDERSTANDING AND RELATED AREAS AT SRI

DONALD E. WALKER

SRI INTERNATIONAL
MENLO PARK, CALIFORNIA

## I INTRODUCTION

SRI International has a long history of research on natural language and on speech. The groups working in these areas were brought together specifically to work on the development of a speech understanding system, but their activities range much more broadly. As a result, SRI is qualified to engage in a variety of projects relating to voice technology for interactive systems applications:

- The design and development of speech understanding systems varying widely in complexity and context of application.
- Research on syntax, semantics, and discourse as they relate to speech recognition and speech understanding systems.
- The integration of practical natural language interface capabilities into systems for speech recognition and voice control.
- Acoustic-phonetic research.
- The development of procedures for speech analysis and speech synthesis.
- The evaluation of the intelligibility and quality of speech, both human and computer-generated, and of communication systems that carry it.
- The identification of properties of speech that contribute to specific qualities, such as "naturalness" and talker identification, and the development of computer procedures for using these properties.
- Studies of the effects on speech of abnormal physiological and psychological states and the development of voice analysis algorithms to detect those effects.

- The conduct of experiments to study the relationships among parameters like the quality of computer speech and computer understanding, the effectiveness of task performance, and the psychological and physiological states of the users.

The technical review of previous work in Section II will concentrate exclusively on our research on speech understanding. It will be followed in Section III by a brief description of our current capabilities for research on speech understanding, speech recognition, and voice control. Section IV will present relevant current research activities; although some of these activities involve text input rather than speech, it would be possible to adapt them to voice control.

## II. TECHNICAL REVIEW OF PREVIOUS WORK

### A. Introduction

From 1971 to 1976, SRI International participated in a major program of research on the analysis of continuous speech by computer sponsored by the Advanced Research Projects Agency of the Department of Defense.* The goal was the development of a speech understanding system capable of engaging a human operator in a natural conversation concerning a specific task domain (see Newell et al., 1973). A rather complex set of specifications defined the parameters more precisely. The program culminated in the demonstration of a system that did meet the target specifications (see Reddy et al., 1976; Medress et al., 1977). However, more important for the future of this technology are developments in the various constituents or sources of knowledge used in the systems--particularly phonetics, phonology, syntax, semantics, and discourse--and in the system architecture necessary for coordinating them efficiently and effectively.

At SRI, we have made signigicant advances in the development both of the components that provide knowledge for use in a speech understanding system and of a framework for coordinating and controlling them. Our work in the ARPA Program was conducted in two phases. During the first phase, we were responsible for the entire system. During the second phase, we worked cooperatively with the System Development Corporation (SDC). For this joint system development effort, SRI provided capabilities for system organization and control, syntax, semantics, and discourse analysis; SDC provided capabilities for signal processing, acoustics, phonetics, and phonology. In this paper, only the SRI work is considered. In the following description, we discuss first our more recent work, since it represents the latest results of our research on

--------

*This research was funded under the following ARPA contracts, all administered through the Army Research Office: DAHCO4-72-C-0009, DAHCO4-75-C-0006, and DAAG29-76-C-0011.

speech understanding, and since we have conducted experiments that have enabled us to provide a partial evaluation of its effectiveness. The subsequent presentation of our earlier efforts considers only the acoustic processing components; it is included to illustrate the research we have done in the speech sciences in the context of speech understanding.

## B. Recent Research on Speech Understanding

### 1. Introduction

Our research on speech understanding has been designed specifically to handle naturally occurring speech, conversations that would take place as a person uses the system on a regular basis as an adjunct to his regular technical activities. A distinctive characteristic of our approach is its emphasis on the relevance of contributions from computational linguistics and artificial intelligence. In processing ordinary conversational dialog, the various sources of acoustic uncertainty combine with the large number of linguistic choices to create an extremely large number of alternative hypotheses that must be considered during the interpretation of an utterance. To control the combinatorial explosion and limit the number of choices that has to be considered, we have introduced sophisticated components for combining information about the structure of English sentences (syntactic knowledge), about the task being considered (semantic knowledge), and about previous utterances in the dialog (dicourse knowledge). To cope with the added complexity of these extra sources of knowledge, we have provided special procedures for coordinating their interactions.

The syntactic component of our speech understanding system is a performance grammar; it describes the syntax of the English occurring in spontaneous dialog rather than the English of edited text. Semantic knowledge about the task domain is encoded in a partitioned semantic network. Partitioning the network allows us, among other things, to represent multiple alternative parses without using excessive storage and to associate syntactic units directly with their semantic counterparts. The discourse component uses the context of the preceding dialog to identify the entities referred to by pronouns and definite noun phrases and to expand incomplete (elliptical) utterances.

Our approach to the coordination of these knowledge sources, and those containing acoustic, phonetic, and phonological information, stresses integration--the process of forming a unified system out of a collection of components--and control--the dynamic direction of the overall activity of the system during the processing of an input utterance. Our approach to integration

- Allows specifying the interactions of information from various sources of knowledge in a procedural representation.

- Provides a means for adjusting the language accepted as input for different tasks without loss of generality.

- Avoids commitment to a particular system control strategy.

Our approach to system control

- Allows processing an input left-to-right, right-to-left, or from the middle out.

- Enables combining top-down, predictive procedures, with bottom-up, data-directed procedures.

- Allows evaluating partial results (phrases) within the larger linguistic contexts (sentences) in which they could be embedded.

A review of the total project is beyond the scope of this paper. After discussing the task domain and presenting an overview of the operation of the system to provide context, we will consider each of the system knowledge sources together with discussions of a facility for language definition that provides the basis for coordinating them and of the executive routines that control them. A brief statement on the results of our experiments with alternative system control strategies also is included.

A more complete statement of this work is contained in our final project report (Walker, 1976). A somewhat expanded description of the language definition system and executive and of the experiments conducted to test them is presented in Paxton (1977; see also Walker and Paxton et al., 1977). The discourse component is treated more fully in Grosz (1977a; see also 1977b for a discussion of the concept of focus). Fikes and Hendrix (1977) summarize the scheme for semantic representation and the procedures for deductive retrieval used in the system. References to other papers are included in the final project report.

## 2. The Task Domain

The domain of discourse for the speech understanding system is defined by a data base of information about the ships of the U.S., Soviet, and British fleets. The system data base contains such characteristics as owner, builder, size, and speed for several hundred ships. Utterances can be formulated that relate to attributes of a particular ship or of ships meeting a certain description; to part-subpart relations between a ship and, for example, its crew; to set membership and kind relationships between various individuals and classes (such as "all ships" and "Are all ships diesels?"). It is possible to specify an object on the basis of its properties ("What country owns the Skate?";

"What American destroyer has a speed of 33 knots?") or of the number of individuals meeting a given description ("How many diesel submarines are owned by the U.S.?"). Queries may be quantified to seek information over classes of individuals ("What is the speed of each American sub?"). Dialog sequences can be processed, with previous utterances serving as context so that pronouns can be used, the referents of determined noun phrases can be identified, and it is not necessary to use complete utterances if the reference is clear ("What is the speed of the Lafayette?"; "The Ethan Allen?"; "Do both ships belong to the U.S.?"; "Are they both submarines?").

### 3. The Operation of the System

When a speaker records an utterance, it is analyzed acoustically and phonetically, and the results are stored in a file. When these data are available, the executive begins to predict words and phrases, guided by the rules for phrase formation in the language definition, and to build up phrases from words that have been identified acoustically in the utterance. When a word is predicted at a specified place in the utterance, alternative phonological forms of that word are mapped onto the acoustic data for that place, and a score indicating the degree of correspondence is returned. As each phrase is constructed, relevant semantic and discourse information is checked, and if appropriate, a semantic network representation of the phrase is developed. When an interpretation for the entire utterance is complete, relevant structures from the semantic model of the domain and from an associated relational data base are processed to identify in semantic network form the content of an appropriate response. This response is then generated either in text form or through the use of a speech synthesizer.

### 4. The Language Definition

The input language is a subset of natural, colloquial English that is suitable for carrying on a dialog between a user and the system regarding information in the data base. The definition of this language consists of a lexicon containing the vocabulary and a set of composition rules for combining words and phrases into larger phrases. This language definition is translated by a definition compiler into an efficient internal representation, which is used by the executive to process an utterance. The lexicon is separated into categories, such as noun and verb, and the words in each category are assigned values for various attributes, such as particular grammatical features and semantic representations. The composition rules are phrase-structure rules augmented by a procedure that is executed whenever the rule constructs a phrase. Information provided by the procedure includes both attributes of the phrase based on the attributes of its constituents, and factors for use in judging the acceptability of the phrase.

An attribute statement may compute values that specify acoustic properties related to the input signal, syntactic properties such as mood (declarative or interrogative) and number (singular or plural), semantic properties such as the semantic network representation of the meaning of the phrase, and discourse properties such as the entity a pronoun refers to. The values of constituent attributes are used in computing the attributes of larger phrases, and the attributes of complete interpretations are used in generating responses.

The factor statements compute acceptability ratings for an instance of the phrase. The factors are non-Boolean; that is, they may assume a wide range of values. As a result, a proposed instance of a phrase is not necessarily simply accepted or rejected; it may be rated as more or less acceptable, depending on a combination of factor values. Like attributes, factors may be acoustic, syntactic, semantic, or discourse related. Acoustic factors reflect how well the words match the actual input; syntactic factors deal with tests like number agreement between various constituents; semantic factors assure that the phrase has a meaning in the task domain; and discourse factors indicate whether a pronoun or definite noun phrase makes sense in the given dialog context. The values of factors are included in a composite score for the phrase. The scores for constituents are combined with the factor scores to produce the scores of larger phrases, and the scores of complete interpretations are used in setting executive priorities.

The attribute and factor statements in the procedural parts of the rules contain specifications for most of the potential interactions among system components. The form of the rules is designed to avoid commitments to particular system control strategies. For example, the rule procedures can be executed with any subset of constituents, so incomplete phrases can be constructed to provide intermediate results, and it is not necessary to acquire constituents in a strictly left-to-right order.

### 5. Syntax

The syntactic knowledge in the system is represented both in the phrase structure part of the language definition rules and in the attribute and factor statements in the procedure part of the rules. Syntax provides computationally inexpensive information about which words or phrases may combine and how well they go together. In testing word or phrase combinations, syntactic information alone often can reject an incorrect phrase without requiring costly semantic and discourse analysis. Factors are used for traditional syntactic tests, such as agreement for person or number, but factors also are used to reduce the scores of unlikely phrases. For example, questions that are negative (e.g., "What submarine doesn't the U.S. own?") are not likely to occur. A factor statement lowers the value for this interpretation but does not eliminate it completely, so that if no better hypothesis can be formed to account for the input utterance, this interpretation will be accepted. Since the language definition system provides the capability for evaluating phrases in context by means of non-Boolean factors, the grammar can be tuned to particular discourse situations and language users simply by adjusting factors that enhance or diminish the acceptability of particular interpretations. It is not necessary to rewrite the language definition for each new domain.

### 6. Semantics

The system's knowledge about the task domain is embodied in a partitioned semantic network. A semantic network consists of a collection of nodes and arcs where each node represents an object (a physical object, situation, event, set, or the like) and each arc represents a binary relation. The network model of the task serves as a foundation on which the structures corresponding to new utterances are built. It is used to assess the feasibility of combining utterance constituents to form larger phrases. And it is a source of information for answering queries, supplemented by a relational data base, which can be accessed directly from the network.

The structure of our semantic networks differs from that of conventional networks in that nodes and arcs are partitioned into spaces. These spaces, playing in networks a role roughly analogous to that played by parentheses in logical notation, group information into bundles that help to condense and organize the network's knowledge. Network partitioning serves a variety of purposes in the speech understanding system:

- Encoding logical connectives and higher-order predicates, especially quantifiers.
- Associating syntactic units with their network images.
- Interrelating new inputs with previous network knowledge while maintaining a definite boundary between the new and the old.
- Simultaneously encoding in one network structure multiple hypotheses concerning alternative incorporations of a given constituent into larger phrases.
- Sharing network representations among competing hypotheses.
- Maintaining intermediate results during the question-answering process.
- Defining hierarchies of local contexts for discourse analysis.

### 7. Discourse

The discourse knowledge in the speech understanding system is used to relate a given utterance (or a portion of it) to the overall dialog context and to entities and structures in the domain. The procedures we have developed are based on systematic studies of dialogs between two people performing some activity together. Contextual influences were found to operate on two different levels in a discourse. The global context--the total discourse and situational setting--provides one set of constraints on the interpretation of an utterance. These constraints are used in identifying the referents of pronouns and definite noun phrases. The second set of constraints is provided by the immediate context of closely preceding utterances. These constraints are used to expand utterance fragments into complete utterances. Since the task domain of the system is data base retrieval, the discourse context is limited to a linear history of preceding interactions. For complex task-oriented dialogs, the linear discourse history can be replaced by a more structured history related to the organization of the task being performed.

### 8. Deduction

Along with the ability to represent entities and their interrelationships in a task domain, it is necessary to reason about them. Thus, the system also contains an inference mechanism for retrieving information from the semantic network. This mechanism serves a dual purpose: (1) during the interpretation of an utterance, it supplies information needed to produce the appropriate semantic structure corresponding to each phrase and to relate it to the dialog context; (2) after an interpretation has been found for a question, it is used to find an answer. This inference capability can retrieve information explicitly stored in the networks, can derive information using general statements, or theorems, in the network, and can invoke user-supplied functions to obtain information from knowledge sources other than the network, such as data files.

### 9. Generation

We also have developed the capability of generating, as a response from the system, an English phrase or sentence that corresponds to a semantic network substructure. This substructure usually is the answer to a question asked by the user. Words and phrases are chosen to express the semantic content; a syntactic frame for their organization is selected; and the response is expressed in text form, although we have sometimes used a commercial speech synthesizer to produce a spoken output.

## 10. Executive

The executive has three main responsibilities:

- It coordinates the work of the other parts of the system calling acoustic processes and applying language definition rules.
- It assigns priorities to the various tasks in the system.
- It organizes hypotheses and results so that information common to alternative hypotheses is shared, avoiding duplication of effort.

When a successful interpretation has been found, the executive invokes the response functions, which produce a reply.

The principal data structure used by the executive is called the parse net. It is a network with two types of nodes: phrases and predictions. Phrases are built from words or from smaller phrases by applying composition rules from the language definition. Phrases can be complete, containing all their constituents, or incomplete, with some or all of their constituents missing. A prediction is for a particular category of phrase associated with a particular location in the utterance. As the interpretation of an utterance progresses, new phrases that have been constructed from existing phrases or from words found in the utterance are added to the parse net. At the same time, new predictions are made as more information is obtained. Thus, as the interpretation process advances, the parse net, which holds intermediate hypotheses and results, grows. A complete root category phrase (usually a sentence) with its attributes and factors constitutes an interpretation of the utterance.

There are two tasks entailed in maintaining and evolving this parse net: the word task and the predict task. The role of the word task is to look for a particular word in a particular location in the utterance. If the acoustic mapper has not been called previously for that word in that location, the word task calls it. If a word is found successfully in the specified location, the word is used to build new phrases. The role of the predict task is to make a prediction for a word or phrase that can help complete an incomplete phrase. Whenever a new constituent is inserted into an incomplete phrase, any adjacent constituents that had been missing can be predicted. New predictions can include predictions for particular words, leading to new instances of calls on the word task.

Establishing the priority of a task begins with determining the score of the phrase involved. The score is computed from the results of the acoustic mapping of any of the words contained in the phrase, from the factor statements for the phrase, and from the scores of the constituents. The score is thus a local, context-free piece of information about how good the phrase is. After the score is determined, the phrase is given a rating that is an estimate of the best score for a phrase of the root (sentence) category that uses the given phrase. The rating for a phrase does depend on the other phrases in which it may be embedded to form a sentence. This rating is then modified depending on the control strategy being used, and the result is the priority of the task to be performed for that phrase.

Both the word and the predict task can work either left-to-right through an input or bidirectionally from words selected at arbitrary positions within an utterance. This ability to add constituents to phrases in any order has made it possible to experiment with a variety of control strategies. Also important for experimental studies is the fact that each task does a limited amount of processing and then stops after scheduling further operations for later. The scheduling does not specify a particular time, but instead gives each operation a certain priority. The operation is performed when its priority is highest. Since the executive sets the task priorities, changing the way these priorities are set alters the overall system strategy.

## 11. Experimental Results

Loss of the computer facility at the System Development Corporation shortly after the system was implemented prevented extensive exercising of the complete system with the acoustic processing components. However, using a simulation of those components, we were able to perform a variety of experiments to analyze the effect of variations in control strategy on system performance. We used an analysis of variance procedure to study four variables:

- To check context or not: use the effects of sentential context based on attribute and factor information in setting priorities versus using only constituent structure information. Context checking should provide more information for setting priorities and should lead to better predictions, but it could prove costly and result in poorer performance.

- To island drive or not: go in both directions from arbitrary starting points in the input versus proceeding strictly left to right from the beginning. Island driving allows interpretations to be built up around words that match well anywhere in the input, but the process is more complex.

- To map all or one: test all the words at once at a given location versus trying them one at a time and delaying further testing when a good match is found. Mapping all at once identifies the best acoustic candidates and reduces the chances of following false paths, but it takes substantially more time.

- To focus or not: assign priorities for tasks focusing on selected alternatives by inhibiting competion versus proceeding each time with the task with the highest score. Focusing prevents frequent switching among alternatives, but it may result in continuing along false paths.

All combinations of the four control-strategy variables were tested on 60 sentences that varied in length, vocabulary, and sentence type.

The results of most interest are those relating to the effects of context checking, that is, using the attribute and factor information. Significant increases in accuracy were found; there was a higher percentage of utterances for which the correct sequence of words was found. Fewer phrases were constructed, so there was less work for the system to do. Rule factors blocked 27% of the attempts on the average; and whereas the average number of phrases constructed over all system configurations was 267, the most accurate system with context checking averaged 158. The percentage of incorrectly identified words was reduced; there was a lower priority for looking at words adjacent to such false alarms than there was for looking at words adjacent to correct words. Finally, the total processing time was reduced, in spite of the extra executive processing required.

These experiments did not provide unequivocal data on how the system would perform with actual rather than simulated acoustic processing components. However, for a lexicon of over 300 words, the most accurate system configuration identified 73% of the utterances. If minor errors that would have no effect on the response of the system are ignored, the figure is increased to 82%. Modifications in the executive alone could increase this latter figure to 90%. Improvements in the acoustic processing components, which the loss of the SDC computer never allowed time to refine, could be expected to increase this figure further. For example, a 7% downward shift in the distribution of scores for words incorrectly accepted by the acoustics would result in a 13% increase in accuracy.

Much more work would be necessary to provide a comprehensive evaluation of our research on speech understanding. However, it is clear that we have produced system control concepts and a set of system components that are well-suited for further research on unconstrained naturally occuring speech.

## C. SRI Research on Acoustic Processing for Speech Understanding

Our earlier work on acoustic processing for speech understanding was conducted in the context of a system design concept that was similar to but simpler than the one described above. The control strategy was exclusively top down; that is, syntactic and semantic information relevant for the current discourse context was used to predict the set of words that could possibly occur at a given place in the utterance. Using data derived from a speech analysis subsystem, a word verification subsystem determined for each proposed word: (1) the confidence that the proposed word did in fact exist at the specified place in the utterance, and, if it could be present, (2) where the word began and ended. The parser, in this version of the system, proceeded through the utterance from left-to-right according to a search strategy that kept track of all possible paths, at any particular moment following the one with the highest priority.

The speech analysis subsystem classified each 10-ms portion of the digitized signal into one of ten classes based on a classification algorithm using digital filter information. The classes were chosen because they would give reliable information in a context-free manner. In addition, a linear predictive coding (LPC) analysis of the voiced intervals provided frequency and bandwidth information for the first five formants. All of the acoustic data in this preprocessing step were stored for each utterance.

The word verification subsystem consisted of a set of algorithms representing the words in the vocabulary. Each such word function was prepared after a detailed examination of acoustic data for that word in selected contexts from a variety of utterances. A word function consisted of a series of Fortran subroutines that used data from a variety of sources: the acoustic preprocessing of the utterance; algorithms for level (volume) detection, formant smoothing, detecting formant discontinuities, fitting formant trajectories, and identifying formant bandwidths; and specially designed digital filters or LPC analyses.

The system that incorporated these acoustic components was not tested extensively, so no conclusions can be made regarding its performance. Of 71 utterances processed by the system, 62% were understood correctly, 10% misunderstood, and 28% not understood at all. We were encouraged by the results of these early efforts, and the experiences influenced our subsequent work.

## III. CURRENT CAPABILITIES FOR SPEECH UNDERSTANDING RESEARCH

### A. Facilities Available

The major computer facility used for our research on speech understanding was a Digital Equipment Corporation PDP/KA-10. It provided time-shared computing capabilities supporting a large variety of programming languages, LISP and Fortran being the ones used most frequently. Currently, SRI has a DEC PDP/KL-10 (System 1090T), which is a larger, faster computer with similar characteristics; it is being used in most of the projects described under Current Research Activities in the following section.

For our early acoustic research, we developed a very powerful interactive speech analysis system. This system, based upon a Vector-General Display controlled by a DEC PDP-15 connected to the PDP/KA-10 computer, allowed scientists to digitize speech, present speech both aurally and visually, edit and mark time series, calculate and display Fourier transforms and LPC analyses of selected portions of speech, calculate and display the results of classification algorithms, plot Formant trajectories, etc. The system was the major tool in the development of the acoustic-phonetic analysis algorithms used in the speech understanding system.

The speech analysis system currently is being upgraded to employ a Hughes Conographics Display controlled by a PDP-11/40 connected over the ARPANET to the PDP/KL-10 computer. The PDP-11 also is connected to an SPS-41 fast array processor that provides real-time calculations of complex speech algorithms such as LPC spectral analysis.

Complementing this system is a PDP-11-controlled psychophysiological laboratory with facilities for digitizing and recording 64 channels of voice and electrophysiological data, including beat-by-beat heart rate, skin conductance response, peripheral pulse volume, respiration rate, and electroencephalographic and electromyographic data from as many as eight subjects simultaneously. We also have a PDP-11-controlled psychophysics laboratory that is used for automated presentation of auditory and/or visual stimuli to subjects and automated recording of responses. Both of these PDP-11 computers are connected to the PDP/KL-10 computer, so that data can be analyzed on the time-shared system.

A Threshold Technology VIP-100 system, which is interfaced to a PDP-11, provides capabilities for isolated word and phrase recognition. We also have a Federal Screw Works VOTRAX ML-I Multi-Lingual Voice System for synthesizing speech.

## B. Personnel

Computational Linguistics and Artificial Intelligence:

Barbara J. Grosz--natural language understanding, discourse analysis, knowledge representation

Gary G. Hendrix--natural language semantics, knowledge representation, semantic network architecture, practical natural language interfaces

Jerry R. Hobbs--text processing, natural language semantics

Gordon S. Novak--question-answering systems, data-base semantics

Ann E. Robinson--language understanding systems, semantic representation and problem solving

Jane J. Robinson--syntax, semantics, phonology, discourse, case and performance grammars, prosodics

Earl D. Sacerdoti--natural language systems for data access, decision aids for command and control

Jonathan Slocum--language generation, semantic network architecture, syntax, semantics, and case systems

Donald E. Walker--language understanding systems, natural language systems for data access, text processing

Speech Sciences:

Richard W. Becker--acoustic-phonetics, speech and speaker recongition by computer, design of large-scale interactive computer systems for speech analysis

Earl J. Craighill--integrated data and voice communication networks, interactive graphic display programming, application of packet radio technology to command and control

Michael H. Hecker--acoustic-phonetics, speech and speaker recognition by computer, forensic applications of speaker identification, effects of pathologies on speech

Fausto Poza--acoustic-phonetics, speech and speaker recognition by computer forensic applications of speaker identification, effects of physiological states on speech

James R. Young--speech and speaker recognition by computer, speech signal analysis and signal processing

## IV. RELEVANT CURRENT RESEARCH ACTIVITIES

### A. Natural Language Understanding Using Text Input

Under ARPA support (Contract DAAG29-76-C-0012), we are providing natural language capabilities in a Navy command and control context. The objective is to develop the technology needed to support a series of increasingly sophisticated systems that provide natural language access to multiple data base management systems over the ARPANET in real time. Each system in the series accepts natural language questions about the data--currently in text form, plans a sequence of appropriate queries to the data base management system to answer each question, determines on which computer to execute the queries, establishes links to those machines over the ARPANET, monitors prosecution of the queries, recovers from certain errors in execution, and prepares a relevant answer to the original question.

Under National Science Foundation support (Grant MCS76-22004), we are developing natural language capabilities for use in intelligent systems that can function as experts, advising and supporting human efforts over a range of problem areas. The objective of the research is to define formally the knowledge necessary for effective communication in natural language between a person and a computer, when they are cooperating on a shared task. Our major emphasis in the project are: (1) the investigation of the structure of dialogs about a task, and (2) the use of the contexts provided by the dialog and the task as aids in understanding utterances. Our activities center on the development of representations for the various kinds of knowledge necessary for understanding utterances and on the development of effective computational procedures for using that knowledge to interpret a sequence of such utterances in a dialog.

A distinctive feature of the project is its concern with understanding the language that occurs in dialogs which take place in a dynamically changing environment. Most other current research either analyzes independent questions or statements within a static environment, as in information retrieval from a computer data base, or considers narratives rather than dialogs, as in story understanding. In contrast, we are interpreting a coherent dialog in relation to an ongoing or previously executed task in which the context can be continually changing. Capabilities are being developed for representing structural features of dialogs and tasks and for dealing explicitly with utterances that relate to past and future, as well as present, and to hypothetical, as well as actual, conditions. Attainment of the goals of this research is essential for the development of intelligent systems that can function as experts, advising and supporting human efforts over a critical range of problems.

### B. Practical Natural Language Interfaces with Text Input

Under SRI Internal Research and Development support, we have been developing and testing LIFER (Hendrix, 1977), a practical system for creating English language interfaces to other computer software (such as data base management systems and expert consultant programs). Its purpose is to make the competence of other computing systems more readily accessible by overcoming the language barriers separating these systems from potential users. Emphasizing human engineering, LIFER has bundled natural language specification and parsing technology into a single package, which includes an automatic facility for handling inputs that do not form complete sentences, a spelling corrector, a grammar editor, and a mechanism that allows even novices, through the use of paraphrase, to extend the language recognized by the system. Offering a range of capabilities that supports both simple and complex interfaces, LIFER allows beginning interface builders to rapidly create workable systems and gives ambitious builders the tools needed to produce powerful and efficient language definitions. Experience with LIFER has shown that for some applications, very comfortable interfaces can be created in a matter of days. The resulting systems are directly usable by such people as business executives, office workers, and military officials whose areas of expertise are outside of computer science. The initial system developed for the ARPA project, referenced above, used LIFER. Other applications provide access to a medical data base and to an interactive photointerpretation system.

### C. Speech-Related Research at SRI

The acoustic facilities at SRI are being used in a variety of research projects. Using the first version of our interactive speech analysis system, we developed a Semi-Automatic Voice Verification System for the Law Enforcement and Administration Agency (Grant NI 71-078-G) that is currently being put into operation. The new speech analysis system, while not yet complete, has already been used in an ARPA project (Contract N00039-76-C-0363) to simulate a Packet Switched Speech Network in a study of the effects of varying system parameters, such as delay and loss of packets, on the efficiency of two-person communication. Under U.S. Government support (Contract 10123-6281770047-7WR) the facilities of the psychophysiological laboratory are being employed to obtain voice and physiological data on 150 subjects to form a data base that can be used to relate speech characteristics to the physiological state of a person in various situations. Within the psychophysics laboratory, we have investigated the effects of phase in human hearing and are currently evaluating the intelligibility and qualtiy of various kinds of machine-processed speech under Defense Communications Agency support (Contract DCA 160-77-C-004).

### D. Practical Uses of Voice Control in Industrial Automation

Under NSF support (currently, Grant APR75-13074), we have been conducting exploratory research into advanced automation. The object of the project is to develop a programmable and adaptable computer-controlled system of manipulators, end-effectors, and contact or non-contact sensors that can be easily trained to perform material handling, inspection, and assembly tasks of the kind that are encountered in industrial settings. The VIP-100 provides voice control to guide a Unimate manipulator in this process. For example, to establish a particular fastening operation the operator, using only spoken words or phrases for control, can train the hand to go through a sequence of positions at several spots in a desired pattern. After training the system in this manner, a single spoken command will cause the system to retrace its sequence of stored actions.

We have just begun a research effort to adapt the parsing techniques of the LIFER system for use with the VIP-100. The resulting prototype system will provide a much more sophisticated capability for responding to complex spoken commands.

## V. REFERENCES

Fikes, Richard E., and Hendrix, Gary G. A Network-Based Knowledge Representation and its Natural Deduction System. Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, Massachusetts, 22-25 August 1977.

Grosz, Barbara J. The Representation and Use of Focus in Dialog Understanding, Ph.D. Dissertation, University of California, Berkeley, California, 1977. (a)

Grosz, Barbara J. The Representation and Use of Focus in a System for Understanding Dialogs. Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, Massachusetts, 22-25 August 1977. (b)

Hendrix, Gary G. Human Engineering for Applied Natural Language Processing. Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, Massachusetts, 22-25 August 1977.

Medress, Mark F., et al. Speech Understanding Systems: Report of a Steering Committee. SIGART Newsletter, April 1977, 62, 4-8.

Newell, Allen Et Al. Soeech Understanding Systems. North-Holland Publishing Company, Amsterdam, 1973.

Paxton, William H. A Framework for Speech Understanding. Ph.D. Dissertation, Stanford University, Stanford, California, 1977.

Reddy, D. Raj. Speech Understanding Systems: Summary Results of the Five-Year Research Effort. Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pennsylvania, September 1976.

Walker, Donald E., (Ed.) Speech Understanding Research. Final Report, Project 4762, Artificial Intelligence Center, Stanford Research Institute, Menlo Park, California, October 1976.

Walker, Donald E., and Paxton, William H., with Grosz, Barbara J., Hendrix, Gary G., Robinson, Ann E., Robinson Jane J., and Slocum, Jonathan. Procedures for Integrating Knowledge in a Speech Understanding System. Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, Massachusetts, 22-25 August 1977. Pp. 36-42.

## BIOGRAPHICAL SKETCH

### Donald E. Walker

Senior Research Linguist
Artificial Intelligence Center
Information Science and Engineering Division

SPECIALIZED PROFESSIONAL COMPETENCE

Computational linguistics; language understanding systems--speech and text: interactive systems for data access through natural language; artificial intelligence strategies for information integration; text processing

REPRESENTATIVE RESEARCH ASSIGNMENTS AT SRI (Since 1971)

Project leader, research on natural language communication with computers for task performance

Project leader, research on integrating tactical information from multiple sources

Project leader, research on speech understanding (ARPA program)

OTHER PROFESSIONAL EXPERIENCE

Head, language and text processing, the MITRE Corporation: Computer-based syntactic analysis procedures for transformational grammars; question-answering systems; personal text file systems

Research affiliate, Linguistics Group, Research Laboratory of Electronics, Massachusetts Institute of Technology

Assistant professor of psychology, Rice University

Visiting assistant professor and research associate, University of Chicago

Research psychologist, Houston Veterans Administration Hospital, and Research associate, Baylor University College of Medicine: Language variability in psychiatric patients

## ACADEMIC BACKGROUND

Ph.D. (1955), University of Chicago

Social Science Research Council Fellow in linguistics, Yale University

## PROFESSIONAL ASSOCIATIONS AND HONORS

American Federation of Information Processing Societies (secretary; member, board of directors); Association for Computational Linguistics (president, vice-president, secretary-treasurer); Association for Computing Machinery (national lecturer); American Society for Information Science (chairman, Special Interest Group on Automated Language Processing); International Federation for Documentation (chairman, Committee on Linguistics in Documentation); International Joint Conferences on Artificial Intelligence (general chairman, program chairman, trustee, secretary-treasurer); Linguistic Society of America

Editorial Boards: Artificial Intelligence, American Journal of Computational Linguistics (managing editor)

Phi Beta Kappa: Sigma Xi

## PUBLICATIONS

More than 25 papers on computational linguistics, artificial intelligence, computer & information science, linguistics, psychology, and anthropology

Editor: _Natural Language in Information Science: Interactive Bibliographic Search; Proceedings of the International Joint Conference on Artificial Intelligence; Information System Science and Technology; Information System Sciences_

(This page intentionally left blank)

D4

[redacted]

N93-72616
S4-32
176340

# VOICE TECHNOLOGY AND BBN

JARED J. WOLF

BOLT BERANEK AND NEWMAN INC.
CAMBRIDGE, MASSACHUSETTS

PRECEDING PAGE BLANK NOT FILMED

## 1. PREVIOUS WORK IN VOICE TECHNOLOGY

Bolt Beranek and Newman Inc. has engaged in research, development, and consulting on a broad spectrum of speech-related problems for over two decades. We have done work in at least the following areas:

- speech signal processing
- automatic speech recognition
- continuous speech understanding
- speaker recognition
- speech compression
- subjective and objective evaluation of speech communication systems
- measurement of the intelligibility and quality of speech when degraded by noise or other masking stimuli
- speech synthesis
- instructional aids for second-language learning and for training of the deaf
- investigation of speech correlates of psychological stress

In addition to these speech-related areas, we also work in experimental psychology, control systems, and human factors engineering, which are often relevant to the proper design and operation of speech systems.

The review of BBN's past and present speech-related projects presented below should not be regarded as delimiting our expertise or research interests. Given our role as an R&D and consulting firm,

PRECEDING PAGE BLANK NOT FILMED

PAGE 64 INTENTIONALLY [redacted]

[redacted]

they represent only specific places where our expertise and interests have intersected with needs of our clients.

## 1.1 Speech Understanding

BBN was a principal participant in the recent five-year Speech Understanding Research (SUR) project, sponsored by the Advanced Research Projects Agency (ARPA) of the Department of Defense. The objective of the SUR project research was to discover, evaluate, and to incorporate into a total system, techniques for using higher level linguistic constraints and advanced signal processing and acoustic-phonetic analysis to determine the best possible interpretation of an unknown speech utterance. These speech understanding systems were to:

> "... accept continuous speech from many cooperative speakers of the General American dialect, in a quiet room over a good quality microphone, allowing slight tuning of the system per speaker, but requiring only natural adaptation by the user, permitting a slightly selected vocabulary of 1ØØØ words, with a highly artificial syntax and a (well defined) task..... tolerating less than 1Ø% semantic error, in a few times real time (on a 1ØØ Mips machine), and be demonstrable in 1976 with a moderate chance of success."

BBN's speech understanding system, called HWIM (for Hear What I Mean), is a powerful research system for exploring alternative control strategies and the effects of different system features. We have used this system to develop some powerful speech understanding algorithms. System components include:

a) A linear predictive coding signal analysis component, which derives smooth spectral parameters, formant and pitch tracks, and other parametric information from the input speech waveform,

b) An acoustic-phonetic recognition component, which segments the acoustic input into a lattice of alternative possible phonetic labelings of the input,

c) An off-line dictionary generation component, which uses within-word and between-word phonological rules to produce word pronunciations expected to be encountered in fluent continuous speech,

d) A fast lexical retrieval component, which can efficiently find words in the vocabulary that match well acoustically with the speech input and which accounts for context-dependent across-word phonological effects,

e) An analysis-by-synthesis word verification component, which can synthesize the expected parametric representation of a hypothesized word (and its context) and compare it with the input parameters,

f) A grammar for interactions with a travel budget management system in natural English using a vocabulary of over 1ØØØ words,

g) A bi-directional parser for ATN grammars, which can parse a sentence from left-to-right, right-to-left, or middle-out,

h) A semantic network knowledge base, which contains general knowledge about trips and places, as well as specific information about planned trips, estimated costs, budgets, expenditures, etc., and

i) A flexible control component, which uses the other components to formulate, evaluate, and extend hypotheses into a complete interpretation of the sentence.

HWIM's speech understanding is set in the context of a travel budget manager's automated assistant, which keeps track of trips taken and planned and the budgets to which trip costs are charged, and it also allows the user to plan new trips. Users may interact with HWIM by speaking sentences from a rather general grammar (over 1ØØØ words, with a high average branching ratio and rejoining paths) forming a subset of natural English. Typical sentences from this task are:

How much is left in the speech understanding budget?
List all trips to California this year.
What is the round-trip fare to Chicago?
Cancel Jerry's trip to the ASA meeting.

At the end of the SUR project in October 1976, HWIM correctly understood about half of its test utterances, spoken by three speakers. (1,4,7-13,16,18,19,23-29)

Continuous speech understanding systems with the capabilities of HWIM and the other ARPA SUR project systems are not yet ready for immediate application, but that was not the goal of the ARPA SUR project. That goal was the development of an advanced technology of speech recognition and understanding. The technology developed during the ARPA SUR project has clear utility in speech recognition and understanding applications that should be practical in the immediate future.

## 1.2 Speech Bandwidth Compression

BBN has been doing research in the speech compression area since 1972, with support from ARPA, and more recently from other sponsors also. BBN has been and is currently involved in developing speech compression systems with a wide range of transmission bit rates, ranging from 75 to 16ØØØ bits/sec, and with different operating conditions such as noisy or high-quality input speech, noisy or noise-free transmission channel, and fixed-rate (synchronous) or variable-rate (asynchronous) transmission. (2,9-13,16,21,22)

The overall goal of the ARPA speech compression research has been to develop linear predictive speech compression (LPC) systems that transmit good quality speech at low data rates. Speech compression techniques developed in this project have been designed for their use in the ARPA Network environment of packet-switched data communications, though they are easily extendible to other communications environments.

Recently developed techniques in linear prediction are used for the analysis and synthesis. We have developed several methods for reducing the redundancy in the speech signal without sacrificing speech quality. Included among these methods are preemphasis of the incoming speech signal, adaptive optimal selection of predictor order, optimal selection and quantization of transmission parameters, variable frame rate transmission, optimal encoding, and improved synthesis methodology. When we incorporated all of these in a floating point simulation of a pitch-excited linear predictive vocoder, we obtained synthesized speech with high quality at average transmission rates as low as 15ØØ bits/sec (21,22). Our more recent results include: development of a new class of stable linear predictive speech analysis methods (12); specifications for an asynchronous or variable data rate linear predictive speech compression system to be implemented by the various ARPA-sponsored sites for real-time speech transmission over the ARPA Network; application of nonlinear spectral warping techniques to either improve speech quality at a given bit rate, or to lower the transmission bit rate at a given speech quality.

One of the major results of the ARPA speech compression project has been to demonstrate real-time speech transmission on the packet-switched ARPA network. BBN participated in the implementation of the SPS-41-based initial system. More recently, a real-time system specified by BBN, transmitting at an average rate of 22ØØ bits/sec, has been implemented on a Floating Point Systems AP-12ØB at Information Sciences Institute. The system will be implemented at BBN on the AP-12ØB we are about to receive.

Our work on speech compression also includes the development of objective procedures for testing the quality of vocoded (or compressed) speech (15,2Ø). Since the objective procedures must be validated against results from subjective listening tests, we also have a program for the subjective evaluation of speech quality. We have explored the perceptual dimensions of speech quality by multidimensional scaling methods (2).

## 1.3 Very-Low Rate Vocoder

An interesting outgrowth of our work in speech understanding, speech compression, and speech synthesis was a project combining phonetic speech transmission system operating at 75 bits per second (14). Based on this pilot project, we have proposed a real time implementation for such a system.

## 1.4 Speech Synthesis by Rule

Our experience in speech synthesis is derived mainly from the research in synthesis-by-rule being carried out by Dennis Klatt at MIT and at BBN (6,7). In our speech understanding system, synthesis played two roles, as a voice response component and as a component of an acoustic-phonetic word verifier, in which a hypothesized word (plus context, if any) was synthesized into an idealized time-varying spectral representation that was then compared against the analyzed utterance itself. In this way, generative acoustic-phonetic knowledge was used to evaluate how well a hypothesized word matched a portion of the utterance (1,4,5). In the phonetic speech transmission system, the receiver used a modification of the synthesis-by-rule program to resynthesize speech from the transmitted values of phoneme identify, duration, and fundamental frequency (14).

## 1.5 Instructional Aids Systems

The instructional aids systems are self-contained computer-based systems for real-time speech analysis and display. A minicomputer receives information about speech-related waveforms via microphones and accelerometers connected to analog and digital preprocessing circuits. Algorithms for analysis and display operate on the data, sometimes under the control of the user, in such a way as to provide concurrent visual and auditory representations of speech sound that may be useful to the user in the modification of his articulation.

The second-language training system is designed to supplement the standard language laboratory. It allows a student to visually compare his efforts with pre-recorded teacher's versions. This system has been evaluated in the context of two language pairs: English speakers learning Chinese and Spanish speakers learning English (3).

The deaf-training system involves a trained teacher working with the student, with the system operating as a tool to enhance their interaction. In this case, attempts have been made to develop displays that are appropriate for use with very young children with severe language limitations as well as profound hearing losses. The prototype system is now being tested at two schools for the deaf (17).

## 1.6 Other Projects

Other projects dealing with voice technology include:

- adapting our variable frame rate speech compression approach to fixed rate transmission operating at 24ØØ bits/sec over a noisy transmission channel,

- ultra-high quality analysis/synthesis of telephone quality speech at 16ØØØ bits/second or less, where the resynthesized speech must be equal in quality to the original input, and

- an investigation of how the psychological state of the user may be reflected in his speech characteristics.

## 2. PRESENT PROJECTS IN VOICE TECHNOLOGY

With one exception, our current research projects in speech processing are continuations of some of the projects described above.

Our work in low rate speech compression continues in the direction of improving the quality of vocoded speech without sacrificing low data transmission rates. Presently under advanced testing is an improved voice source model incorporating both periodic and noise components, which largely eliminates the "buzziness" often associated with vocoded speech. We will also be bringing into real-time vocoder implementation many of the quality improvement techniques already demonstrated in our floating point vocoder simulations. We also expect to be starting work on high-quality speech synthesis of the type required for a very-low-rate phonetic vocoder system.

Also continuing are the projects on:

- variable-to-fixed rate transmission over a noisy channel
- ultra-high quality analysis/synthesis at a 16 kbit rate
- vocal indicators of the speaker's psychological state.

One new project, not mentioned above, is to develop a processing system to improve the intelligibility of speech that has been corrupted by wideband noise.

## 3. ANTICIPATED CAPABILITIES IN VOICE TECHNOLOGY

### 3.1 Staff

With its experience in a wide variety of projects dealing with voice processing, BBN numbers among its staff many with training and experience in the field. In 1977, 11 full-time scientists and 3 regular consultants are engaged in voice technology research and development, almost all of these with advanced degrees. We expect to maintain at least this level of staffing in the foreseeable future. BBN's Information Sciences Division, within which our speech projects are based, numbers over 1ØØ scientists from a broad variety of fields, particularly computer science, artificial intelligence, computational linguistics, electrical engineering, and the behavioral sciences.

### 3.2 Facilities

The BBN Research Computer Center (RCC) has four DEC PDP-1Ø's and one DECsystem-20. Three of the PDP-1Ø's run TENEX, a virtual memory time sharing system developed by BBN. The other PDP-1Ø and the DECsystem-2Ø run TOPS-2Ø, a DEC supported time sharing system based on TENEX. Much of the speech processing work not requiring real-time processing is carried out on the KL 1Ø/9ØT system which runs TOPS-2Ø. All of the program libraries used in the speech and signal processing are runnable on both TENEX and TOPS-2Ø.

BBN's Speech Processing Laboratory contains equipment for speech signal acquisition, display, editing, storage, and playback, and it provides a facility for advanced real-time speech processing systems research and development. It currently includes a DEC PDP-11/4Ø, a Signal Processing Systems Inc. SPS-41 signal processor (including dual A/D and D/A converters), and an Imlac PDS-1 graphics display processor. Delivery of a Floating Point Systems Inc. AP-12ØB array processor is scheduled for the beginning of calendar 1978; this addition will substantially enhance our real-time processing capabilities. The PDP-11/system is connected to the ARPANET, which is used for data and program transfers to and from the RCC or any other site on the ARPANET, and for packet speech experiments for our continuing speech compression projects. The Laboratory also contains audio equipment for producing, manipulating, and recording audio signals.

## 4. REFERENCES

(1) Cook, C., and R. Schwartz (1977), "Advanced Acoustic Techniques in Automatic Speech Understanding," IEEE International Conference on Acoustics, Speech and Signal Processing, Hartford, CT, 1977, pp. 663-666.

(2) Huggins, A.W.F., R. Viswanathan, and J. Makhoul (1977) "Quality Ratings of LPC Vocoders: Effects of Number of Poles, Quantization, and Frame Rate," Proc. 1977 IEEE International Conf. on Acoustics, Speech and Signal Processing, Hartford, CT, 1977, pp. 413-416.

(3) Kalikow, D.N., and J.A. Swets (1972) "Experiments with Computer-Controlled Displays in Second-Language Learning," IEEE Trans. Audio Electroacoust., AU-2∅, 23-27.

(4) Klatt, D.H. (1975) "Word Verification in a Speech Understanding System," in D.R. Reddy (ed.), Speech Recognition: Invited Papers Presented at the IEEE Symposium, Academic Press, 321-341 (see also Bolt Beranek and Newman Inc., Report No. 3∅82, Cambridge, MA).

(5) Klatt, D.H. (1975) "The Design of a Machine for Speech Understanding," in Speech Communication, Vol. 3, G. Fant (ed.), Halsted Press, pp. 277-289.

(6) Klatt, D.H. (1976) "Structure of a Phonological Rule Component for a Synthesis-by-Rule Program," IEEE Trans. Acoustics, Speech, and Signal Processing ASSP-24, pp. 391-398.

(7) Klatt, D.H., C.C. Cook and W.A. Woods (1975) "PCOMPILER -- A Language for Stating Phonological and Phonetic Rules," Bolt Beranek and Newman Inc., Report No. 3∅8∅, Cambridge, MA, pp. 18-23.

(8) Klatt, D.H. and K.N. Stevens (1973) "On the Automatic Recognition of Continuous Speech: Implications of a Spectrogram-Reading Experiment," IEEE Tran. on Audio and Electroacoustics AU-21, 21∅-217.

(9) Makhoul, J. (1973) "Spectral Analysis of Speech by Linear Prediction," IEEE Trans. Audio and Electroacoustics, AU-21, 3, pp. 14∅-148, June 1973.

(10) Makhoul, J. (1974) "Linear Prediction vs. Analysis-by-Synthesis," Speech Communication Seminar, Stockholm, Sweden, Vol. 1, pp. 35-43, Aug. 1974.

(11) Makhoul, J. (1975) "Linear Prediction in Automatic Speech Recognition," in Speech Recognition: invited papers presented at the IEEE Symposium, D.R. Reddy (ed.), New York: Academic Press, pp. 183-22∅, 1975.

(12) Makhoul, J. (1975) "Linear Prediction: A Tutorial Review," invited paper, IEEE Proceedings, special issure on Digital Signal Processing Vol. 63, No. 4, pp. 561-58∅, April 1975.

(13) Makhoul, J. (1975) "Spectral Linear Prediction: Properties and Applications," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, No. 3, pp. 283-296, June 1975.

(14) Makhoul, J., R. Schwartz, C. Cook, and D. Klatt (1977) "A Feasibility Study of Very Low Rate Speech Compression Systems," BBN Report No. 35Ø8, Bolt Beranek and Newman Inc., Cambridge, Mass., February 1977.

(15) Makhoul, J., R. Viswanathan and W. Russell (1976) "A Framework for the Objective Evaluation of Vocoder Speech Quality," IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, pp. 1Ø3-1Ø6, April 1976.

(16) Makhoul, J., and J. Wolf (1972) "Linear Prediction and the Spectral Analysis of Speech," Report No. 23Ø4, Bolt Beranek and Newman Inc., Cambridge, Mass., Aug. 1972.

(17) Nickerson, R.S., D.N. Kalikow, and K.N. Stevens (1974) "A Computer-based System of Speech-Training Aids for the Deaf," BBN Report No. 29Ø1. Abbreviated version published in AFIPS Conference Proceedings, 43, pp. 125-126.

(18) Schwartz, R., and J. Makhoul (1974) "Where the Phonemes Are: Dealing with Ambiguity in Acoustic-Phonetic Recognition," IEEE Symposium on Speech Recognition, Contributed Papers, Carnegie-Mellon Univ., Pittsburgh, PA, pp. 85-88, April 1974. Also in IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, No. 1, pp. 5Ø-53, Feb. 1975.

(19) Schwartz, R. and V. Zue (1976) "Acoustic-Phonetic Recognition in BBN SPEECHLIS," IEEE International Conference on Acoustics, Speech and Signal Processing, April 12-14, 1976, Philadelphia, 1976, pp. 21-24.

(2Ø) Viswanathan, R., J. Makhoul and W. Russell (1976) "Towards Perceptually Consistent Measures of Spectral Distance," IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, pp. 485-488, April 1976.

(21) Viswanathan, R., J. Makhoul, and R. Wicke (1977) "The Application of a Functional Perceptual Model of Speech to Variable-rate LPC Systems," Proc. 1977 International Conference on Acoustics, Speech and Signal Processing, Hartford, CT, 1977.

(22) Viswanathan, R., and J. Makhoul (1975) "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans.

Acoustics, Speech and Signal Processing, Vol. ASSP-23, No. 3, pp. 3Ø9-321, June 1975.

(23) Wolf, J.J. (1976) "Knowledge, Hypotheses, and Control in the HWIM Speech Understanding System," Conf. Record, 1976 Joint Workshop on Pattern Recognition and Artificial Intelligence (IEEE Catalog No. 76CH1169-2C), Hyannis, MA, June 1-3, pp. 113-125.

(24) Wolf, J.J. (1976) "Speech Recognition and Understanding," in K.S. Fu (ed.), Digital Pattern Recognition, Springer-Verlag, Berlin, Heidelberg, New York.

(25) Wolf, J.J. (1977) "HWIM, A Natural Language Speech Understander," Conference Record, 1977 IEEE Conference on Decision and Control, New Orleans, La., 7-9 December 1977.

(26) Wolf, J.J. and W.A. Woods (1977) "The HWIM Speech Understanding System," Conference Record, IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, Ct, pp. 784-787, May 1977.

(27) Woods, W.A. (1977) "Shortfall and Density Scoring Strategies for Speech Understanding Control," 1977 Int'l Joint Conference on Artificial Intelligence, MIT, Cambridge, Aug. 22-25, 1977.

(28) Woods, W.A., et al. (1974) "Natural Communication with Computers: Speech Understanding Research BBN," BBN Report No. 2976, Vol. I, Bolt Beranek and Newman Inc., Cambridge, Mass., Dec. 1974.

(29) Woods, W.A., et al. (1976) "Speech Understanding Systems: Final Report," BBN Report No. 3438, Vols. I-V, Bolt Beranek and Newman Inc., Cambridge, Mass., December 1976.

## BIOGRAPHICAL SKETCH

### Jared J. Wolf

B.E.E. (summa cum laude), Union College, Schenectady, N.Y., 1965; S.M., Ph.D. (Electrical Engineering), Massachusetts Institute of Technology, 1967, 1969.

As a graduate student, staff member, and Research Associate in the Speech Communications group at M.I.T., Dr. Wolf did research on techniques of speech analysis and speaker recognition. He was a Leverhulme Postdoctoral Fellow at the Department of Psychology, University of Edinburgh, Scotland, from 1970 to 1971. Since 1971, he has been a Senior Scientist at Bolt Beranek and Newman Inc., Cambridge, MA, where he has primarily been concerned with the development of speech understanding systems, particularly in the areas of signal processing, phonological rules, and control strategy.

Dr. Wolf is the author of many papers and reports dealing with signal processing, speech analysis, speaker recognition, and speech recognition and understanding.

(This page intentionally left blank)

# SPERRY UNIVAC SPEECH COMMUNICATIONS TECHNOLOGY

MARK F. MEDRESS

SPERRY UNIVAC DEFENSE SYSTEMS DIVISION
ST. PAUL, MINNESOTA

## INTRODUCTION

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

## CONTINUOUS SPEECH RECOGNITION

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

### The Recognition System

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

[redacted]

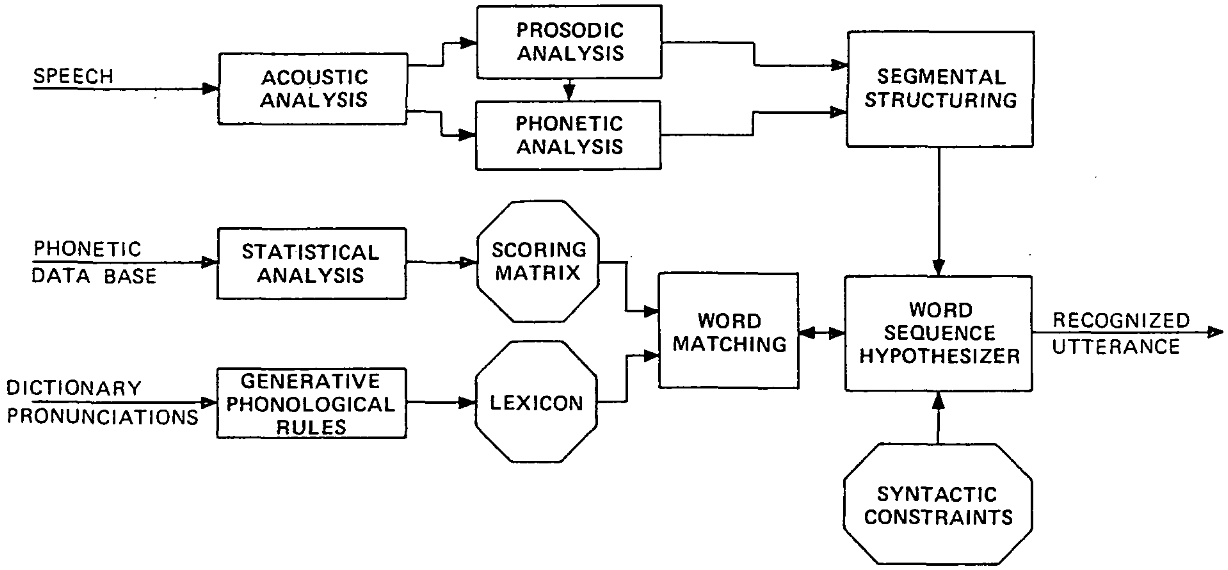


Figure 1. The Sperry Univac Continuous Speech Recognition System

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

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

## A Recognition Example

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

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

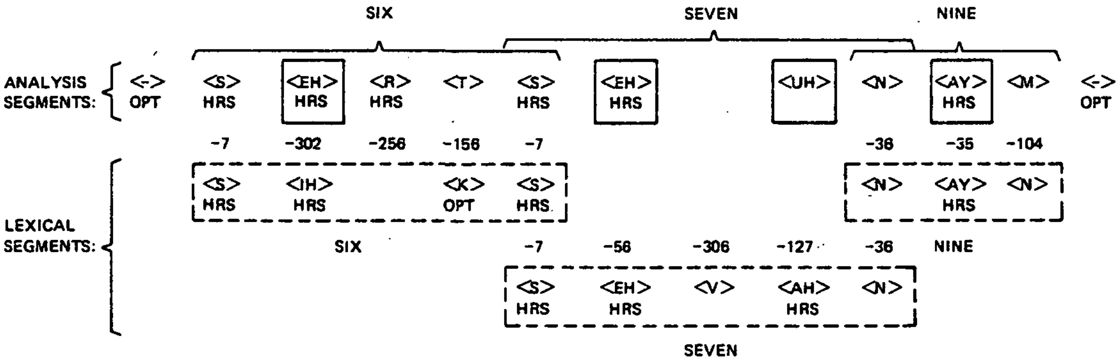


Figure 2. Recognition of the Phrase "Six Seven Nine"

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

## The Recognition Data Base

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

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

ALPHANUMERIC SEQUENCES

- Vocabulary size = 36
- Average branching factor = 9.4

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

DATA MANAGEMENT COMMANDS

- Vocabulary size = 64
- Average branching factor = 6.3

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

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

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

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

Recognition Performance and Future Development

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

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

ALPHANUMERIC SEQUENCES

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

DATA MANAGEMENT COMMANDS

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

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

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

## OTHER SPEECH COMMUNICATIONS PROJECTS

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

### Word Spotting

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

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

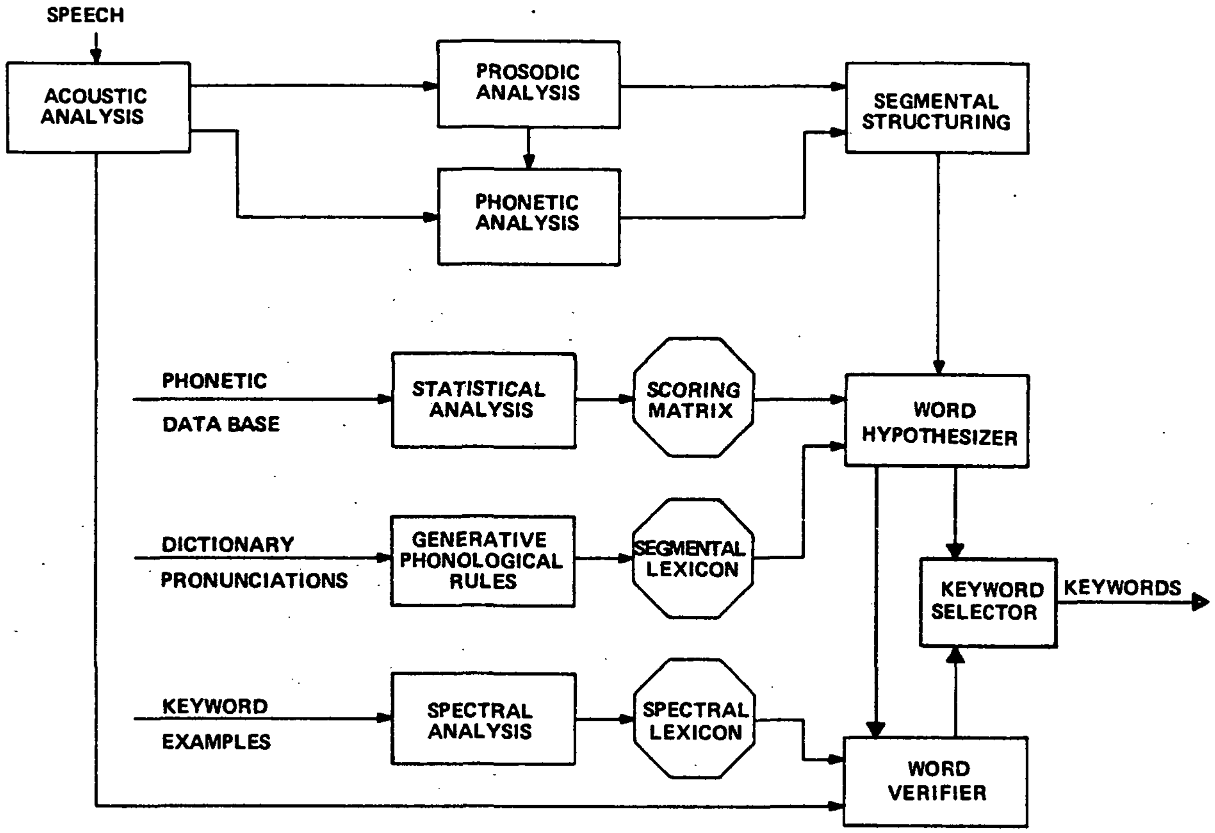


Figure 5. The Sperry Univac Word Spotting System

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

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

## Prosodic Research

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

## Voice Response

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

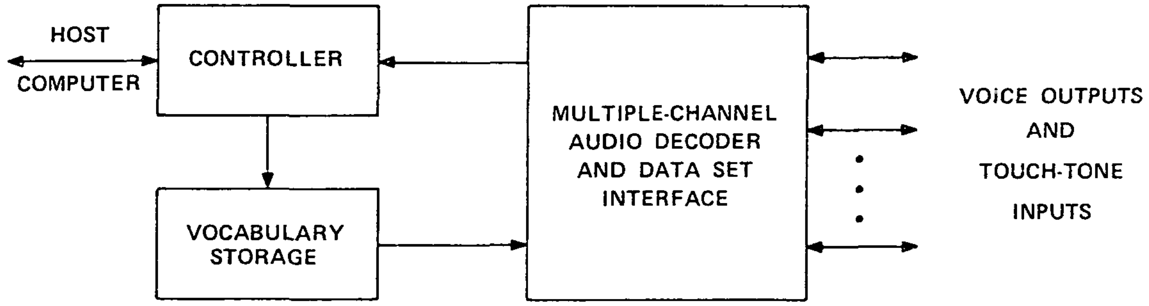


Figure 6. The Sperry Univac Voice Response Unit

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

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

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

## RESOURCES FOR SPEECH TECHNOLOGY DEVELOPMENT

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

### Personnel

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

### Facilities

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

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

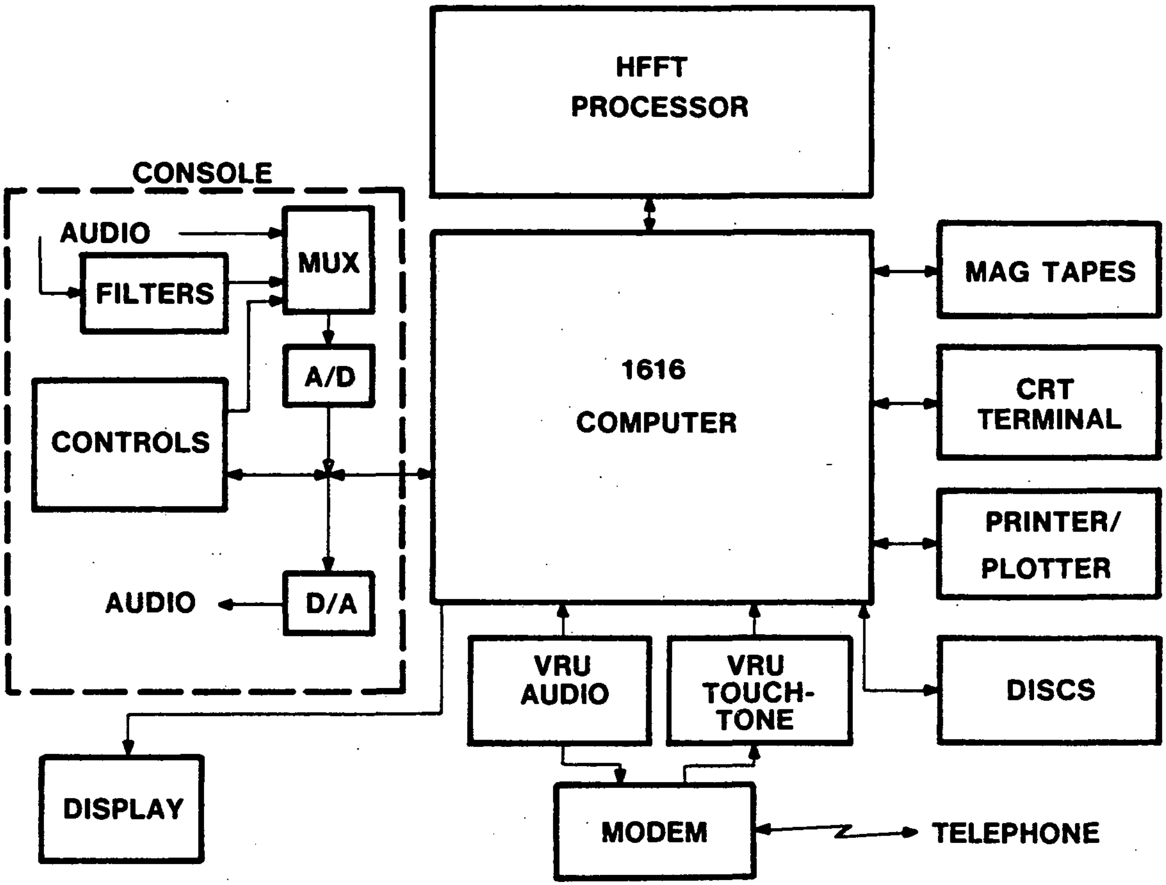


Figure 7. The Sperry Univac Speech Research Facility

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

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

## SUMMARY

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

## REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## BIOGRAPHICAL SKETCH

### Mark F. Medress

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

## DISCUSSION

## Dr. Mark Medress

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

C-2

(This page intentionally left blank)

D6

N93-72618

[redacted]

56-32
176342

# VOICE INPUT/OUTPUT CAPABILITIES AT PERCEPTION TECHNOLOGY CORPORATION

LEON A. FERBER

PERCEPTION TECHNOLOGY CORPORATION
WINCHESTER, MASSACHUSETTS

PRECEDING PAGE BLANK NOT FILMED

Perception Technology Corporation was founded in 1969, and began at that time to engage in speech research based upon a Theory of Speech Perception previously advanced by its founder and president, Dr. Huseyin Yilmaz. Since that time, PTC has undertaken and successfully performed a number of research and development programs in speech for various government agencies. As a result of this experience and the backgrounds of PTC personnel, high level capabilities exist in a number of areas related to speech perception.

The Theory of Speech Perception as proposed by Dr. Yilmaz has undergone expansion and refinement over the years, and has been the basis of the speech research effort at PTC. Phenomena predicted by the theory have been verified experimentally, and recognition equipments emulating the human perceptual capability have been constructed. Arising from this work, recognition algorithms and methods have been developed for speaker-independent recognition, recognition of connected speech, and spotting of specific words in unrestricted context. This background has also taken PTC into the voice response field. We have studied both the waveform and spectral natures of speech, and have gained insight into the human facility of speech communication.

The dominant goal of the work at PTC has been the development of effective speech recognition systems. Upon founding of the company, effort was immediately begun on the first PTC recognizer. When completed in 1970 this machine was capable of speaker independent recognition of the digits with a 98% accuracy. Work has continued both under PTC and government sponsorship to expand the utility of this basic system in areas of connected speech, keyword recognition, increased vocabulary, and speaker acceptance. The present capability as recently reported is a recognition accuracy of 99% on a 20 word vocabulary by 50 speakers. An accuracy of 97% has also been realized by a recognizer for connected digits. A more detailed description of capabilities and the performance of the speech recognition systems at PTC is given in the facility section of this paper.

PRECEDING PAGE BLANK NOT FILMED

PAGE 94 INTENTIONALLY BLANK

[redacted]

The above discussion is a sample of the capabilities of PTC to carry out programs of research and equipment development. This experience qualifies PTC to undertake related tasks through ability of its personnel to grasp and comprehend high-level concepts such as speech perception, and also through their abilities in implementation of these concepts by computer.

## FACILITY DESCRIPTION

Perception Technology Corporation maintains two fully equipped laboratories and a production area. One laboratory is equipped with all the standard and special purpose instruments for R&D in the areas of signal and speech processing, and with instruments and components for breadboarding and testing digital and linear electronic circuits and systems. Another laboratory is equipped for general research in perception and audio perception in particular. It includes equipment to generate speech or to manipulate audio signals to generate a wide range of stimuli required for perception studies in speech. The production area is equipped for assembly of circuit boards and for light manufacturing.

The computer facility configuration shown in Figure 1 is a block diagram showing the major hardware components of the various speech recognition systems. The main system is based on the PDP 11-70 computer operating under RSX-11M. This system is used for software development and for non real-time speech recognition. Most programs are written in FORTRAN IV Plus, evaluated and optimized before conversion to machine language for real-time operation. At the present time this procedure applies only to PDP 11 compatible software. In FY 78 we are planning to have the 11-70 emulate PDP8 and Z80 instructions so that software development for all of PTC's Voice I/O systems can be performed under the main operating system.

There are four additional speech recognition systems, three of which are shown in Figure 2. Two of these are fully operational; the others are under development.

Figure 2a shows the hardware configuration of an on-line data entry system that is planned for FY 78. It is based on software developed for the recognition of digits and control words spoken in connected strings. These programs are now being converted from FORTRAN to assembly language for real-time operation. The system will combine other modes of data entry, such as a digitizing tablet and a CRT, with speech recognition. The system will recognize the English digits spoken in connected strings of random length, and a set of 15 control words.

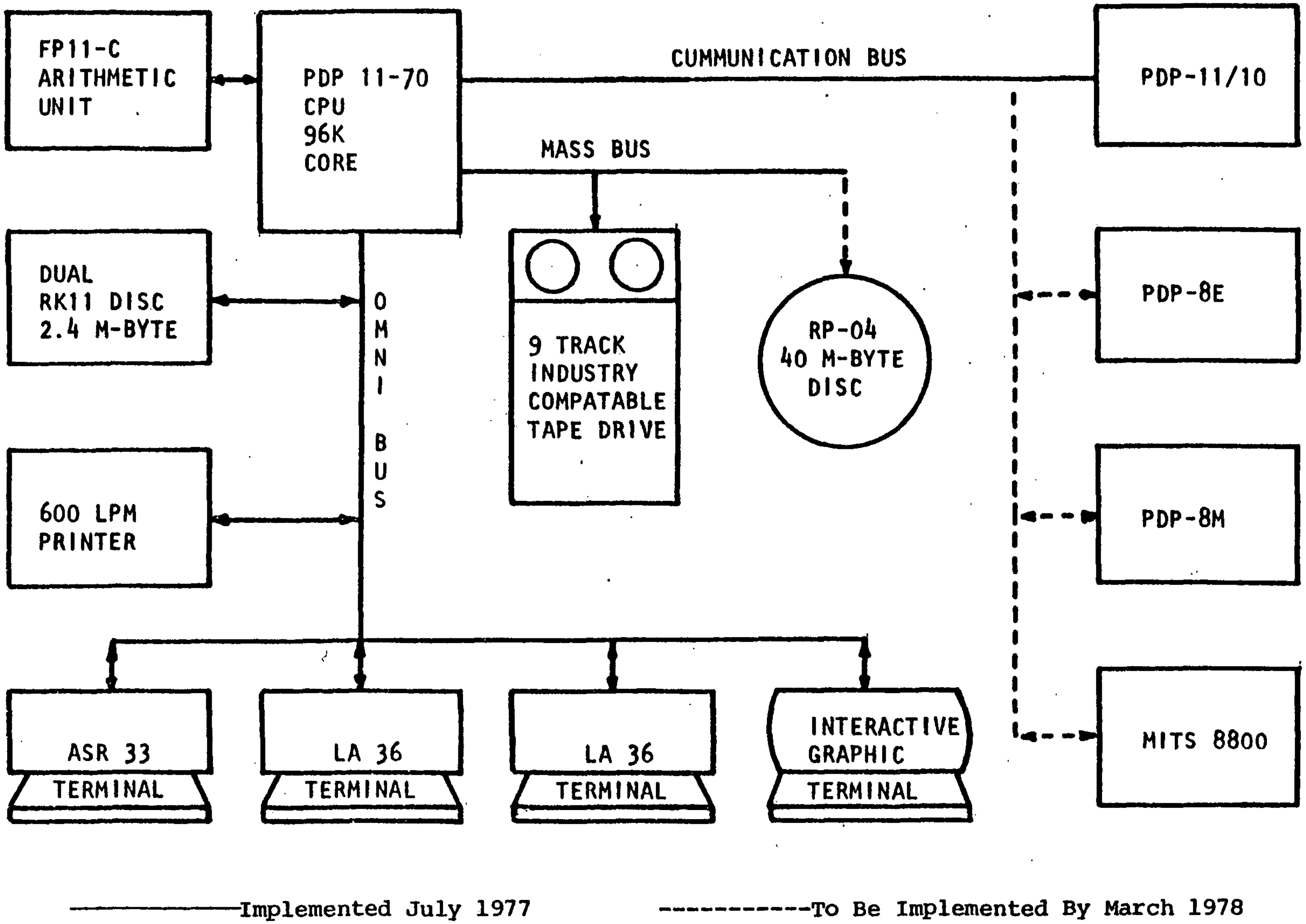


Figure 1. A Block Diagram of The PTC Computer Facility

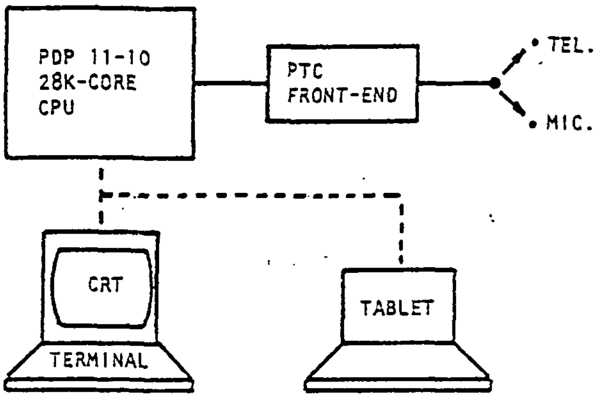


FIG. 2a. Hardware configuration of the real-time, connected speech, recognition system.

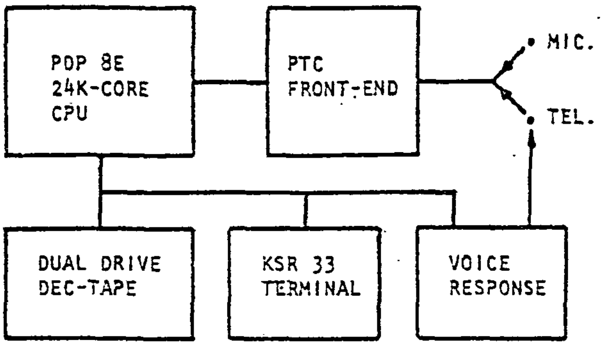


FIG. 2b. Hardware configuration of the Voice I/O on-line system, used for product demonstrations.

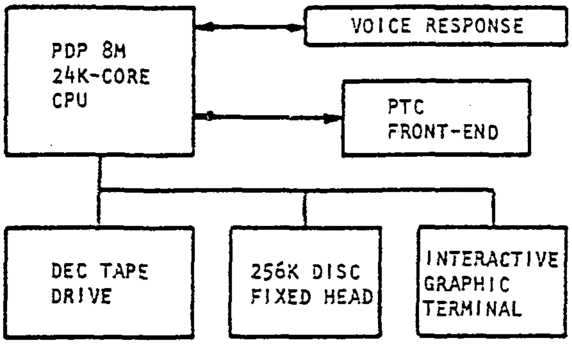


FIG. 2c. Hardware configuration of the Voice I/O non real-time system, used for product software development and connected speech recognition.

Figure 2

The system shown in Figure 2b is a system used for demonstration and evaluation of word recognition. The system is capable of recognizing a syntax-free vocabulary of 30 words spoken in a discrete manner. It is a general-purpose recognizer containing many modes of operation and training schemes. In the speaker independent mode, the vocabulary consists of the 10 digits plus 6 control words. In the trained mode, the vocabulary can be 20-30 words depending on the number of syllables per word. The training has two basic modes of operation, direct training and adaptation. For some applications the two can be combined for increased utility. The direct training consists of repetition of the vocabulary words in sequence or in random fashion using a 32 character alphanumeric display for prompting. In the adaptation mode the system must first be trained for a certain vocabulary, but subsequent speakers use only a few words to get the system adapted to their speech. This system operates with telephone or microphone inputs. The telephone operation is not yet fully interactive; the voice response portion does not yet have a large enough vocabulary for remote prompting and communication.

In FY 78 we are planning to implement the basic recognition portion of the above system on a microprocessor. At the present time we have some of the software operational on an 8080 based development system. The microprocessor based system is expected to be operational by July 1978.

The system shown in Figure 2c is a development system for PDP8 based software. It is also used for testing of "connected speech" recognition and word spotting. The system operates off-line, non real-time and performs recognition on connected digits. Performance tests on this system using constraint-free speech, spoken in random length digit sequences, resulted in an overall recognition accuracy of 97%. This test was done under laboratory conditions using 25 male speakers and results were reported in a technical report No. RADC-TR-76-273.

PTC also maintains a laboratory for general research in perception, and audio perception in particular. The set-up includes equipment to generate speech or to manipulate audio signals for the generation of a wide range of stimuli used in the study of speech perception. This set-up utilizes a PDP8L processor with several software packages. These programs, together with special purpose hardware have been used to implement the following systems:

- An adaptive time compression system for maximizing intelligibility of sped-up speech.

- A digital speech waveform processor with the necessary flexibility for the study and manipulation of signals in the time domain. This system is also used for synthesizing speech and to generate the data base for the voice response unit.

- A pitch-independent display unit for speech training for the handicapped, based on a color-speech analogy.

## SCIENTIFIC & TECHNICAL STAFF

The staff at Perception Technology Corporation consists of seven full-time scientists and engineers with extensive experience in the fields of speech recognition, speech synthesis, speaker authentication and language identification. Other employees include hardware and software engineers with a wealth of experience in system design, circuit design and computer programming. They are augmented by part-time technicians to aid in construction and testing of circuits and systems.

Scientific consultants to and directors of Perception Technology Corporation include: Professor Roman Jakobson of MIT and Harvard University, Professor Harry Levinson of Harvard University and Professor Philip Morse of MIT.

The following pages contain condensed resumes of key company personnel. The information given is pertinent to the fields of research which the company is presently pursuing and does not reflect their overall experience or their achievements in other areas.

### HUSEYIN YILMAZ

Dr. Yilmaz recieved B.S. and M.S. degrees in electrical engineering from the Technical University of Istanbul in 1950 and 1951. In 1952, he enrolled as a doctoral candidate at the Massachusetts Institute of Technology and became a research assistant in physics. He received the Ph.D. in theoretical physics in 1954.

From 1954-56, he was a member of the physics department at the Stevens Institute of Technology and in 1956 became a staff member of the National Research Council of Canada. He joined Sylvania Electric Products in 1957, as an engineering specialist pursuing research with emphasis in the fields of atomic physics, theory of relativity, and color perception.

In 1961, Dr. Yilmaz published a mathematical theory of color perception based on adaptive postulates derived from the Darwinian theory of evolution. More recently, he has generalized this theory to embrace other sense perceptions, including the perception of the residue pitch of the human ear, the perception of speech and psychophysics of sensory organization in audio-visual perceptions.

In the spring of 1962, Dr. Yilmaz joined the Research and Development Division of the Arthur D. Little organization of Cambridge, Mass., becoming a member of the Senior Research Staff and a Staff Consultant. During the years of 1962-64, he was also a Research Associate in the Department of Biology at M.I.T.; a guest, for two months in 1964, of the Institute for Perception Research, Eindhoven, Netherlands; and, in 1965-66, a Visiting Professor (full) in Electrical Engineering at M.I.T.

Currently he is concentrating in the fields of speaker-independent recognition of speech, the psychophysical laws, and the problems of audio-visual perception in general. He has also a new statistical approach to quantum field theory which was published in 1969. This work aims at removing field theory divergences by introducing statistical constraints without violating any of the fundamental principles of physics.

As president and principal investigator at Perception Technology Corporation, Dr. Yilmaz follows a highly interdisciplinary approach and tries to join sophisticated ideas and theories with practical engineering applications.

1. "Psychophysics and Pattern Interactions", Models for the Perception of Speech and Visual Form. (Proceedings of a Symposium. Sponsored by the Data Sciences Laboratory, Air Force Cambridge Research Laboratories, Boston, Mass., Nov. 11-14, 1964), Weiant Wathen-Dunn, ed. Cambridge & London: M.I.T. Press, 1967.

2. "On the Pitch of the Residue", Report No. 41, Institute for Perception Research, Eindhoven, Netherlands, 1964.

3. "On Speech Perception", Report No. 42, Ibid.

4. A Program of Research Directed Toward the Efficient and Accurate Recognition of Human Speech. (I). Prepared for the National Aeronautics & Space Administration, Electronics Research Center, Cambridge, Mass. Cambridge: Arthur D. Little, Inc., Dec. 14, 1966, p. 64.

5. "Speech Perception--I", (Vowels), Bull. Math. Biophysics, 29, Dec. 1967.

6. "A Theory of Speech Perception--II", (Consonants), Bull. Math. Biophysics, 30, Sept. 1968.

7. "A Real-Time, Small Vocabulary, Connected-Word Speech Recognition System" (H. Yilmaz, et.al.) Final Report, Contract No. F30602-72-C-0083, 1972.

8. "Perceptual Continous Speech Recognition" (H. Yilmaz, et.al.) Final Report, Contract No. F30602-74-C-0061, March, 1974.

9. "Automatic Speaker Adaptation" (H. Yilmaz, et.al.) Final Report, Contract No. F30602-75-R-0130, July, 1976.

Dr. Yilmaz has given many invited lectures in the U.S. and abroad on speech and color perception. He is the author of numerous internal reports on word spotting and speech recognition published by various government agencies. In addition, he published two books and more than 50 papers and articles in general relativity and psychophysics.

LEON A. FERBER

Mr. Ferber received his B.S. degree in Electrical Engineering from Northeastern University, Boston, Massachusetts in 1969.

Currently, Mr. Ferber is Vice President of Perception Technology Corporation in charge of basic research and product development. He is involved in the design of the company's line of Voice Input/ Output products and the implementation of computer based systems for industrial control and material handling. His administrative duties include marketing of Voice Input/Output equipment and contract administration.

Mr. Ferber joined Perception Technology Corporation as an Electrical Engineer to design the digital and analog circuits that went into the construction of the company's first speech recognition system. Subsequently, he was in charge of the design and construction of audio instruments for internal use.

During the years 1967-69, Mr. Ferber worked for Digital Equipment Corporation, Maynard, Mass. His work included design and release to production of circuits for automatic memory test systems, interfacing peripheral equipment to the PDP-8 line of small computers and design of display systems.

1. "A Three Parameter Speech Display", Proceedings of the 1972 International Conference on Speech Communication and Processing, Newton, Mass., April 24-26, 1972.

2. "Speech Perception" Final Report, Real Time, Context Free, Connected Speech Recognizer, Contract No. F30602-74-C-0061, April, 1975.

JAMES SHAO

Dr. Shao received his B.S. degree in Electrical Engineering in 1959 and his M.S. degree in Solid State Physics in 1961, all from the University of Birmingham, England. He received his Ph.D. degree in Physics in 1971 from Massachusetts Institute of Technology.

Presently, Dr. Shao is in charge of development in the area of speech recognition and is the project director on a program to develop a "word spotting" system. His interests are in the areas of speech signal processing and speaker transformation. He also participates actively in the development of computer software necessary for the realization of these processes.

In 1975, Dr. Shao directed the development of a recognizer for unconnected speech. This effort resulted in a product known as PTC VE200.

In 1974, Dr. Shao joined Perception Technology Corporation as a staff scientist to apply symbolic manipulation to the solution of problems in theoretical and applied physics. He participated in the PTC Gravity Research Program and contributed to the study of detection and generation of gravity waves.

From 1972 to the present, Dr. Shao has been a consultant to ERDA at Los Alamos Scientific Laboratory, Los Alamos, New Mexico. He is engaged in the development of software for the Heavy Nucleus Research Program at the Laboratory.

From 1965 to 1968, Dr. Shao was employed by Arthur D. Little, Inc., Cambridge, Mass. he carried out development work on solid state devices and he was in charge of the experiments in their speech research program. During this period, he and Dr. Yilmaz explicitly showed the analogy between color perception and speech perception.

MICHAEL H. BRILL

Dr. Brill joined Perception Technology Corporation in 1977. As a staff scientist he is responsible for the application of speech perception theories in the area of "word spotting", and "connected speech" recognition. His present work includes: Development of feature selection algorithms, application of probability theory and statistics to speech data base generation.

Dr. Brill received his Ph.D. degree in Physics from Syracuse University in 1974; his thesis was "Color Vision: an Evolutionary Approach". He received a M.S. degree in Physics from Syracuse University in 1971 and a B.A. degree in Physics and English from Case Western Reserve University in 1969.

In the period from 1974-77 Dr. Brill was a Post-Doctoral Fellow at M.I.T. working with Professor J. Y. Lettvin on the psychophysics and neurophysiology of the visual system. His work included: computer simulation of information processing in the human visual system, impulse propagation in nerve fibers, and studies of perceptual invariants. He also taught courses and presented lectures on color and vision.

In 1972 Dr. Brill was with the United States Air Force as a 2nd Lieutenant at the IRAP Division, Rome Air Development Center. He monitored contracts on machine recognition of speech and contributed to in-house research on speaker recognition.

## HENRY G. KELLETT

Mr. Kellett joined Perception Technology in 1971. He was previously Manager of Acoustic Applications at Peripheral Sciences Inc., of Norristown, Pennsylvania, and has worked as a Senior Research and Development Engineering in Speech Recognition at Philco-Ford and Sperry Rand.

Currently, he is a staff scientist contributing to research and development on government sponsored programs in speech recognition based upon a theory of speech perception and its practical application.

In his present position, he has supervised and contributed to contracts for the National Security Agency and Rome Air Development Center. He has previously been responsible for the design and construction of Speech Recognition equipment at Philco-Ford and Peripheral Sciences Inc.

Mr. Kellett received his education in Electrical Engineering at the University of New Hampshire and the University of Pennsylvania, and holds a B.S. degree in Electrical Engineering.

1. "A New Time Domain Analysis Technique for Speech Recognition", Proceedings of the 1972 International Conference on Speech Communication and Processing, Newton, Mass., April 24-26, 1972.

2. "Experimental, Limited Vocabulary, Speech Recognizer", (Co-author), IEEE Transactions on Audio and Electroacoustics, Vol. AU-15, No. 3, September 1967.

3. "Experimental Speech Recognizer for Limited Word Input", Electronic Communicator, Vol. 2, No. 6, Nov./Dec. 1967.

4. Co-author of numerous technical reports for the National Security Agency, and Rome Air Development Center.

DON DEVITT

Mr. DeVitt joined Perception Technology Corporation in 1977 to take over system development and software operations on the RSX-11 operating system. Presently Mr. DeVitt is working on the conversion of PTC's product software from the PDP-8E to a Z-80 based microprocessor. His objective is to construct a low cost, self-adaptive real time word recognizer.

During 1976, while at Tufts University graduate school, Mr. DeVitt worked with Perception Technology on a voice response system. This system, named the BT-2 Voice Output Terminal, later became a part of PTC's product line.

Mr. DeVitt holds a B.S. degree in Electrical Engineering and a M.S. degree in Computer Science. He received his degrees in 1975 and 1977, respectively, from Tufts University.

Prior to joining PTC, Mr. DeVitt worked for First Data Corporation developing software for interactive graphics and signal processing.

## SUMMARY

Recognition methods of connected and continuous speech have been developed by PTC through stratified processing techniques. The smaller, phoneme and syllable, elements are first recognized, then sequences of these are next applied to the large, word and phrase, recognition tasks. This method may be described as a time-warping procedure by which input speech may be recognized even though exact time correspondence does not occur and word boundaries do not correspond with any stored reference data. The methods used in the identification and classification of the phonetic elements are based on a spacial representation corresponding to a perceptual space in which talker and channel transformation are performed. The details of this method are presented in numerous reports that are referenced in the biographical section of this paper. Because of its generality, this method is directly applicable to the implementation of a word identification system. We view all acoustic level speech recognition machines as word spotting systems with appropriate application-oriented constraints. For example, by applying a forced decision threshold and constructing a reference data set for one cooperative speaker, our most general system is reduced to the simplest speech recognizer.

At the present, our main effort is concentrated in the area of word recognition in natural speech. This encompasses two areas of application, keyword spotting and data entry. The keyword spotting

effort is supported by the U.S. Government under contract DAABO-3-75-C-0438. The work in the field of "natural speech" data entry is supported partially by contract No. F30602-77-C-0168 and partially by internal funding. The keyword identification system is targeted as a feasibility study to demonstrate the effectiveness of such a system to perform in a non-cooperative, unknown speaker environment. It is being implemented on a large minicomputer in FORTRAN IV+ and is expected to run in 2-3 times real time. Final evaluation is expected late in FY 78. The data entry system is being implemented on a minicomputer and will operate on-line in real time. A laboratory prototype is expected to be operational in FY 78, and will use speech in combination with other means of data entry. The vocabulary consists of the English digits and command words which may be spoken in connected strings of random length. A similar system operating in an off-line mode was demonstrated at PTC late in FY 76.

## BIOGRAPHICAL SKETCH

### Leon A. Ferber

Leon A. Ferber is Vice President of Perception Technology Corporation. He is responsible for the development, application and marketing of voice input and output systems.

Mr. Ferber received his B.S. degree in Electrical Engineering from Northeastern University, Boston, Massachusetts in 1969. He joined Perception Technology Corporation in 1969 and designed the company's first word recognition system and numerous speech training equipment for the deaf. Since 1972 he has been project manager of continuing government and internal R&D effort in speech recognition.

During the years 1967-1969 Mr. Ferber worked for Digital Equipment Corporation, designing automatic test systems and graphic displays.

omit
to
p122

# SESSION II

DR. ROBERT BREAUX

NAVAL TRAINING EQUIPMENT CENTER, ORLANDO, FLORIDA

This session presents some of the other applications of speech technology. The first session presented a great deal about artificial intelligence. We heard the terms "man/machine interaction", "command and control systems". These terms, we found, mean different things to different researchers. Yesterday's presentations showed that speech is, in fact, a natural communication channel for the interaction of intelligent entities, a human and a machine. But there was some confusion, I think, yesterday. Those talks could have left the impression that the immediate widespread application of speech understanding must wait for the solution of some significant problem. I will have to agree with that. Before we use speech as an artificial intelligence channel we do have some more work to do. But I also must add that there are commercial firms selling speech products to an ever-growing market. These products are marketed as a way for a company to reduce cost, or to increase productivity among it's people.

Since this market is continually expanding, something must be working in the field of automated speech. So let's shift gears now, and see what these systems are about. Yesterday, we were in low gear, and rightly so. We must have a firm foundation of the potential for automated speech technology. And in low gear yesterday we saw some very powerful potentials. Today, let's shift to drive. We will take a look at how and why commercial off-the-shelf products are being used. But let's also keep in mind that when we shift to drive, we don't want our shiny new technology running away with us, whisking us off to applications for which the technology is not ready. To avoid that, those of you representing government agencies wanting to implement automated speech technology should begin your planning with an analysis of the application. Determine first the extent of an artificial intelligence requirement that you have and this can serve as a measure of how to proceed in your application. One of our efforts at the Naval Training Equipment Center's Human Factors Laboratory, where I am employed as a research psychologist, is an effort for the application of automated performance measurement technology to training.

(This page intentionally left blank)

omit

# LABORATORY DEMONSTRATION OF COMPUTER SPEECH RECOGNITION IN TRAINING[1]

DR. ROBERT BREAUX[2]

NAVAL TRAINING EQUIPMENT CENTER
ORLANDO, FLORIDA

## INTRODUCTION

### Background

The Naval Training Equipment Center's Human Factors Laboratory seeks to identify and measure those behaviors which, when improved through training, result in superior performance on the job. Thus, the laboratory seeks to combine new technology developments with current advances in learning/training theory and techniques.

One such technology development is computer speech recognition. The advantage brought to training by this technology is the capability to objectively measure speech behavior. Now, traditional training techniques for jobs which are primarily speech in nature require someone who can listen to what is being said. Otherwise, no measure of the speech behavior is possible. In the U.S. Navy, jobs which are primarily speech in nature include the Ground Controlled Approach (GCA) and Air Intercept (AIC) controllers, as well as the Landing Signal Officer for carrier operations, various Naval Flight Officer positions such as the Radar Intercept Officer, and the Officer of the Deck in ships operations. In addition to the requirement of having an instructor listen to the speech behavior, training in these situations often requires another person to cause changes in the environment which correspond to the trainee's commands. For the GCA and AIC tasks, this takes the form of "pseudo" pilots who "fly" a simulated aircraft target. This 2:1 ratio of support personnel to trainee results in a relatively high training cost.

Previous studies have demonstrated that in analogous situations, it has been possible to achieve savings of manpower and training time while gaining a uniform, high-quality student output by introducing automated adaptive instruction. This advanced technology, if applied to GCA controller training, would bring in its standard benefits such as objective performance measurement and complete individualized instruction.

[1] This paper was presented, in part, at the Tenth Naval Training Equipment Center/Industry Conference, 16 November 1977, Orlando, Florida, and published in the proceedings of that conference.

[2] The opinions expressed here are those of the author and do not necessarily reflect the official policy of the United States Navy.

Moreover, for GCA controller students, a more fully automated system could provide greater realism in the performance of "aircraft" under control by accessing directly the computer model of aircraft dynamics rather than relying on the undetermined skills of a variety of pseudo-pilots. Additionally, the rapid processing of an automated system would make possible extrinsic feedback of task performance to the trainee in real-time.

But in order to realize an automated adaptive training system, it is essential that, in addition to values of overall system performance, some relevant aspect of the trainee's activity, in this case his speech behavior, be accessible to the performance measurement subsystem. At this point, our technology review suggested that the state of the art in machine understanding of speech could furnish the means for direct entry of a trainee's advisories. For some whose acquaintance with this possibility is limited to the science fiction of film, television and print media, the response might be "Of course! Why not?" Those more familiar with the problem might say, "Not yet!" The reality is that while computer understanding of continuous unrestricted speech, without pretraining, by any individual who approaches, is still a long way off, there exists today a capability for machine recognition of isolated utterances drawn from a small set of possible phrases. The computer in this case must be pretrained on the language set with speech samples for each individual speaker.

## Automated Adaptive Instruction

Automated adaptive training has a number of advantages over the more traditional approaches to training. Automation of training relieves the instructor of busywork chores such as equipment setup and bookkeeping. He is thus free to use his time counseling students in his role as training manager. In adding the adaptive component, efficiency is increased with more training per unit time. Individualized instruction, with its self-paced nature maintains the motivation of the trainee. Objective scoring is potentially more consistent than subjective ratings. Uniformity can be maintained in the proficiency level of the end product, the trainee. But, tasks requiring verbal commands have thus far been unamenable to automated adaptive training techniques. Traditionally, performance measurement of verbal commands has required subjective ratings. This has effectively eliminated the potential development of individualized, automated, self-paced curricula for training of the aforementioned Landing Signal Officer, the Air Intercept Officer, the Ground Controlled Approach Controller, and others. Computer speech recognition of human speech offers an alternative to subjective performance measurement by providing a basis of objectively evaluating verbal commands. The current state of the art has allowed such applications as automated baggage handling at Chicago's O'Hare airport. A more sophisticated recognition system is required for training, however. To that end, the Naval Air Systems Command and the Advanced Research Projects Agency have supported the Naval Training Equipment Center Human Factors Laboratory in efforts to establish design guidelines for training systems which combine automated adaptive training technologies with computer speech recognition technology. The particular application chosen is the Precision Approach Radar (PAR) phase of the GCA.

## TRAINING REQUIREMENTS

### The GCA Application

The task of the GCA Controller is to issue advisories to aircraft on the basis of information from a radar indicator containing both azimuth (course) and elevation (glidepath) capabilities. The aircraft target projected on the elevation portion of the indicator is mentally divided into sections by the controller. This is because the radio terminology (R/T) for glidepath is defined in terms of these sections. Thus, at any one point in time, one and only one advisory is correct. Conversely, each advisory means one thing and only one thing. This tightly defined R/T is perfect for application of objective performance measurement. The drawback, of course, is that performance is verbal and has thus far required subjective ratings. In addition, the time required for human judgment results in inefficient performance measurement. The instructor cannot catch all the mistakes when there are many.

### Needs and Objectives

The major behavioral objective of current GCA training is to develop the skill to observe the trend of a target and correctly anticipate the corrections needed to provide a safe approach. The standard R/T is designed to provide medium to carry out this objective, and GCA training exposes the student to as many approaches as possible so that the trainee may develop a high level of fluency with his R/T.

The primary need to fulfill its objective is for GCA training to teach the skill of extrapolation. A controller must recognize as quickly as possible what the pilot's skill is. He must recognize what the wind is doing to the aircraft heading. Then he must integrate this with the type aircraft to determine what advisories to issue.

### Advanced Technology

The major behavioral objectives, then, can more efficiently be achieved through the application of computer speech recognition technology, and thereby the application of advanced training technologies. This is because with objective assessment of what the controller is saying, objective performance meausrement is possible, and thus we have the capability of individualized instruction. The use of simulated environmental conditions allows the development of a syllabus of graduated conceptual complexity. The integration of these components results in an automated, self-paced, individualized, adaptive training system.

The job of the instructor now becomes one of training manager. His experience and skill may be exploited to its fullest. The training system can provide support in introducing the student to the R/T. The instructor can scan the progress of each student and provide counseling to those who need it. Simple error feedback is provided by the training system. Only the instructor can provide human to human counseling for specific needs, and the training system provides more time for this valuable counseling.

## TRAINING SYSTEM OVERVIEW

A training system for the GCA controller was determined to require four subsystems, speech understanding, pilot/aircraft model, performance measurement, and a syllabus. The speech understanding subsystem was developed around the VIP-100 purchased by the Naval Training Equipment Center from Threshold, Inc., Cinnaminson, New Jersey.

Three major constraints are imposed by this system. Each user must pretrain the phrases. Recognition does not take place for random, individual words, only predefined phrases. Each phrase is repeated a number of times and a Reference Array is formed representing the "average" way this speaker voices this particular phrase. Thus, the second constraint is that there must be a small number of phrases (about 50) which are to be recognized. If performance is to be evaluated based upon proper R/T, each phrase must be defined. The third constraint, due to performance measurement requirements, is that there be no ambiguous phrases -- right or wrong depending strictly on who the instructor is. Technically, the GCA application appears to be conformable to these constraints.

To achieve high fidelity, simulation makes use of various math models: The model of the controller is at the focal point of all other models, and serves to provide criteria to the performance measurement system. A model of the aircraft and pilot allows for variation in the complexity of situations presented to the student. The principle being used here is that exposure of a student to certain typical situations will allow him to generalize this experience to real world situations. The pilot model allows for systematic presentation of various skill levels of pilots. In addition, the equations used in modeling the pilot and aircraft responses also allow for introduction of various wind components. The adaptive variables, pilot skill, aircraft characteristics, and wind components are combined systematically to produce a syllabus graduated in problem complexity. As the skill of the trainee increases, he is allowed to attempt more complex problems.

Since the score is determined by the performance measurement system, the heart of scoring is the model controller. As it often happens, what constitutes "the" model controller is a matter of some discussion among GCA instructors. Thus for automated training applications, one must determine the concepts which are definable, such as how to compute a turn, and leave other concepts to be developed by the instructor-student apprentice relationship.

## RESULTS

### The Problem of Novelty

In an attempt to verify the recognition algorithms, naive adult males were employed as subjects. It was soon discovered that probability of correct recognition was as low as 50 percent in the beginning and phrases had to be retrained to increase recognition reliability. It was hypothesized that the novelty of "talking to a machine" was a significant factor in the low-recognition reliability. If this initial novelty could be reduced, it was thought, reliability would also increase. Four adult males and four adult females were used to compare an introduction method vs a no introduction method. The introduction group was given R/T practice, saying the GCA phrases as they would later in an actual prompted run. The model controller was utilized to anticipate for the subject an optimum response every four seconds. This prompt was presented graphically on the display, as the aircraft made the approach. The subject spoke the phrase, then both the prompt and the understood phrases were saved for later printout. The no introduction group, on the other hand, was not given practice. Each group then made reference phrases. Reliability data was collected using the procedures described above for R/T practice. A Chi-square value was computed from a 2 x 2 contingency table of frequency of runs in which no recognition errors occurred vs frequency in which one or more errors occured, and whether there had been practice on the phrases vs no practice prior to making the voice reference patterns. It was found that $X^2(1) = 3.12$, $p<.10$ indicating a relationship. A correlation was computed for the groups vs the number of different phrases which were not recognized on a run with $R = -.33$, $p<.10$, indicating a tendency for fewer errors with pre-practice at the task. Conclusion: Better recognition is achieved when the R/T is voiced consistently and unemotionally.

### Training System Evaluation

Twelve recruits were used form the Recruit Training Command, Orlando, who were in their last few weeks and, therefore, were privileged with liberty on the weekend. Each had received assignment to the Navy's Air Traffic Control (ATC) School. Each subject was interviewed for willingness to participate in an "experiment" during liberty hours concerning ATC, and each was informed that for their time they would be paid. Each subject expressed a desire to become an air controlman.

Each subject was issued at the interview those portions of the programmed instruction booklets normally used by the ATC School relating to the Precision Approach Radar (PAR) phase of GCA, and was requested to complete the material prior to arrival at the lab. Each subject was exposed in the lab to approximately three hours of "introduction". During this time the system collected and validated the voice pattern of the subject for each of the PAR phrases. During the between-run intervals, audio recordings were played which explained and reviewed the PAR R/T. Recognition accuracy by the system on the final run of each subject ranged from 81.5% correct to 98.5% with an average of 94.1% correct recognition.

Subjects were then exposed to "free" runs in which they had complete control over the aircraft. It was found that recognition accuracy suffered during the first few runs. The change from a system which fully prompted the subject on the R/T to a full scoring system which required the subject to initiate all R/T, resulted in a noticeable change in the voicings of the R/T. Hesitation, repetition, and corrections were made which, of course, is not within the capability of the speech system to accurately reocgnize. R/T voicing improved with practice, however.

## Subsequent School Evaluation

The ATC School was informed of which persons had been exposed to the lab PAR system. Eight of the original 12 subjects completed the 14 week school. Four dropped for "various academic and non academic" reasons, and were therefore dropped from further analysis. During school PAR training which followed exposure to the lab system by about 14 weeks, the subjects' average performance was equal to the school average. A product moment correlation was computed for final score at the school vs complexity level achieved on the lab system. The position correlation R = .78, p<.05) indicates that better performance on the lab system was related to higher scores at the school. School instructors reported better than average voicings of the R/T by the subjects exposed to the lab PAR system.

The conclusion drawn was that the lab PAR system taught skills similar to those required at the ATC school and, further, that the use of computer speech recognition can be combined with advanced automated training technology to produce an automated training system for the PAR portion of GCA training. Procurement is underway for an experimental prototype system to be installed and evaluated at the ATC school itself.

## Where From Here

The technology requirements which follow are based on projections for the next three to five years for proposed applications of automated computer speech recognition in training. The single most important need is off-the-shelf hardware (e.g., isolated word recognition (IWR) hardware) with software for a limited continuous speech recognition (LCSR) capability. This must have real-time operation with a vocabulary size of 50-100 words. Since training must assume some degree of naivety on the part of the human speaker, training requires a capability to recognize what was said rather than what was meant. Thus, syntax and grammars, which aid processing of the acoustical signal, can in fact be detrimental to training.

Let's consider an example of LCSR and its impact for training. In the GCA approach, a common error is for the trainee to use the word glideslope rather than the correct term glidepath. Now, IWR systems recognize the entire phrase "slightly above glidepath" as one word. So it is seldom that the error is caught when glideslope is used instead of glidepath. With the LCSR capability, however, such errors could be routinely detected. Further, use of syntax as an aid in "understanding" what was meant by the trainee when he erroneously substituted glideslope for glidepath would result in failure to detect that error.

Speaker independence is popular today as a goal for computer speech technology. However, in the training environment the need exists for recognition of speakers from a large cross-section of the population. In fact, there are foreign nationals being trained by some Navy schools. Therefore, emphasis in the training area is for systems which can recognize highly varied speakers, including English speakers whose native language is not English. The IWR system, with its requirement for speaker pretraining, appears to be sufficiently developed to meet this need, particularly if LCSR were included.

Other technology requirements in the training area are reduced hardware costs, less critical microphone placement, and recognition in a noisy environment. Of course, cost is always a factor in any procurement activity. Microphone placement becomes important when the goal of the training system within which the speech hardware operates is a goal of total automated training. The less critical the mike placement, the more inexperienced the user can be. Finally, noise levels cannot always be reduced, as in flight deck operations. With greater noise tolerance, however, greater application could be made for speech recognition. One such example is simulation of flight deck operations for training the Landing Signal Officer.

## SUMMARY

A system was described which provided a laboratory evaluation of the feasibility of the use of computer speech recognition in training. Results of the evaluation indicate that training can be enhanced and manpower costs reduced by a careful integration of advanced training technology with off-the-shelf computer speech recognition hardware which is enhanced with software algorithms designed for a specific vocabulary set. The need was indicated for further research and development via and experimental prototype system to be installed at the Navy's Air Traffic Control School.

## REFERENCES

Breaux, R. and Grady, M.W. "The Voice Data Collection Program - A Generalized Research Tool for Studies in Speech Recognition." In Proceedings of the Ninth NAVTRAEQUIPCEN/Industry Conference, Technical Report: NAVTRAEQUIPCEN IH-276, Orlando, Florida, Naval Training Equipment Center, November 1976.

Breaux, R. and Goldstein, I. "Developments of Machine Speech Understanding for Automated Instructional Systems." In Proceedings of the Eighth NAVTRAEQUIPCEN/Industry Conference, Orlando, Florida, Naval Training Equipment Center, November 1975.

Goldstein, I., Norman D.A., et al. "Ears for Automated Instructional Systems; Why Try?" In Proceedings of the Seventh NAVTRA-EQUIPCEN IH-240, Orlando, Florida, Naval Training Equipment Center, November 1974.

## ACKNOWLEDGEMENT

The design and implementation of the vast bulk of the software for the Speech Understanding Subsystem and the Performance Measurement Subsystem were done by M.W. Grady, M.J. Barkovic and R.M. Barnhart of Logicon, Inc., San Diego, California. J.P. Charles and L.D. Egan were, successively, Project Manager for Logicon under NAVTRAEQUIPCEN Contract N61339-74-C-0048. Design and implementation of the automated instructor critique software was performed at the Human Factors Laboratory by the author.

## BIOGRAPHICAL SKETCH

### Dr. Robert Breaux

Dr. Robert Breaux received his Ph.D. in experimental psychology from Texas Technical University in 1974. He is a Research Psychologist in the Human Factors Laboratory at the Naval Training Equipment Center. He has an interest in application of the theoretical advances from the psychological laboratory to the classroom situation. Publications and papers include computer application for statistics, basic learning research, concept learning math models, and learning strategies. He is an instrument rated commercial pilot, and a certified flight instructor.

## DISCUSSION

## Dr. Robert Breaux

Q: Roland Paine, Systems Control: You mentioned recognizing words, but on this particular training application it seems emotions and the way he controls his voice is very important as well. Have you addressed that issue at all?

A: That's correct. The disc jockey-like voicings are very important to instructor controllers. One of the points that they like about the isolated word recognition systems and the requirement to create voice reference patterns was to require the trainee to speak almost in a monotone, but more importantly, very consistently. Always say the same thing the same way. If a pilot is coming in with icing on his wings and low fuel, he is excited enough. The controller doesn't need to get excited. We need somebody who is calm and cool. We can simulate situations like that to teach the controller how to handle it. There is a potential that with speech technology's requirement to speak very consistently, the instructors feel that there is the potential to improve that portion of training which is concerned with the training of the RT, the Radio Terminology. The students tend to mimic their instructors a great deal, which means they try to go as fast as they can, be very smooth and suave, etc. The instructors really want them to learn the basics right now. You can develop your own technique later. So in that sense, that's one good point about the isolated word recognition systems,

In addition, there is one problem that is very significant in training to me that is different from the problem in the operations area. And that is related, in a way, to syntax and grammar (this is addressed in the paper, by the way). In the training situation, we have a branching factor equal to the vocabulary size because we need to diagnose what the trainee's problem is. He's not an expert in the situation as a pilot would be. We are not talking about having the trainee saying whatever he wants to say and if the system understands him, make the airplane do that. Although that might be a good application in the operational area, it's not in training. We want to teach him to speak the correct phrases. So a speech understanding system that tried to "hear what I mean", may loose a potential to diagnose what the trainee's weakness is at that point in training and, thereby, loose the potential to determine what sort of situation the trainee may need next. A connected word speech system which could pick out each of the words would be helpful in that sense. Does that answer your question? Any others?

Q: George Doddington, Texas Instruments: Here is the situation. When you are training the controller to do a function where he receives data from a computer and gives data back to the computer, he receives data through a visual display and gives it back to the computer which digests it, recognizes the word and passes it on to the pilot, it seems like an interesting possibility for total automation in this case, where you replace the controller with a computer and the computer then needs to speak to the pilot. What to you think about that idea?

A: Great idea, once you let me describe it this way: Any of you who are pilots realize that you don't trust controllers very much, much less a computer. And even though you might fly a hands off approach on an ACL system, an Automatic Carrier Landing system, you don't fly very far hands off. You're out there ready to grab it. Yes, that's true, and most of the people, a lot of the management-type people who come through our lab, whose job is not concerned necessarily with training or R&D, often make the comment, that gee, what do you need the controller for. And it's certainly a reasonable approach.

Q: George Doddignton, TI: I guess what I am asking is: Is this being considered, are there any programs, have there been any programs, what are the problems? If I were a pilot, I think that I would probably trust the computer more than I would a human, in all seriousness.

A: I won't fly with you. No, I'm kidding. What else can I say? It's a good point.

Q: Wayne Lee, SCRL: If the student is, in fact, going to mimic the instructor and part of his instruction comes from the machine, what quality of speech might be heard. I would't want him to mimic the Votrax we heard.

A: That's a good point. That's been brought up by the instructor controllers themselves.

Q: Wayne Lee: Wouldn't it be very reasonable to just have prerecorded speech that is plugged together and that becomes output?

A: That's a potential that we are considering in the prototype. We'd like to look at a number of ways. As I mentioned the other day, a prototype is a system on which we'll be doing research. I think that was a good point yesterday. We have yet to come out of the lab really. We're going to be in a training situation, but it's going to be a controlled situation and we'd like to look at a number of variables. This again is an R&D effort and when it comes time to procure an operational trainer, if that time comes, then these points should certainly be taken into account, I would think.

Q: Ed Huff, NASA Ames: I don't recall if you mentioned it. What is the language size that you were dealing with and in the course of training, what has been your experience with recognition accuracy? Has that fallen off or improved? And finally, what happens if the recognizer doesn't work properly?

A: First question is vocabulary size, and we are working with a 44 phrase vocabulary. Second question was reocgnition accuracy. Recognition accuracy ranged from about 89 to 97 percent. The third question, in the laboratory version, when I was doing some of the work, I would play an audio tape recorder for part of the time. When the system, the isolated word recognition system, did not understand what was being said, I could replay the audio tape and let the trainee hear what he was saying. In the prototype device we will automate that particular function as well. Essentially, it's a situation which the trainee is trying to learn a number of tasks simultaneously. We hope with advanced training technology that we can reduce these tasks in a small step procedure so that these sort of things don't all hit him at once, and that he won't have trouble voicing his RT. In some situations there are a number of things he must learn all at the same time, not only what to say, but when to say it. He may know exactly what's happening, he's learned that well, and he's just fishing for his RT. He can't think of what to say, and he says "six miles to glide path". You know, little things like this that the system, of course, doesn't recognize. The trainee has the concept; he's fishing for his RT. There are a number of training problems associated with this that are very, very intriguing to me, and that's one of them. Hopefully, we can address some of that in the prototype device. Any other questions?

Q: Dr. Raj Reddy, Carnegie-Mellon University: I have a general comment to make. Those of us who are in artificial intelligence research are constantly faced up to this question of replacing human beings with machines. I think that's a very poor use of words and some of us get carried away with our own enthusiasm. In the long run, I think the way to view this, the use of a computer in general as an intelligent instrument as we better understand how we can encode more and more of the routine knowledge that an Air Traffic Controller or anyone brings to bear on the problem more of that knowledge can be put into the computer so that the person there can use this facility to do the more important planning and other type of tasks. So the thing we should be talking about is intelligent instruments that would aid all of us whether you are a doctor, an engineer, or whether you're a scientist, in doing your job better, to augment your own intellect. I think that's the way we should think of the use of the computers rather than replacement of a human being by the computer. And I get very sensitive, because those of us who work in the field never think about artificial intelligence as a panacea which will do away with the human beings.

A: That's a good point, and I guess I'm sensitive to it in a way too. And the reason is that we tend to be more intellectual at times than, say another group of people. Keep in mind that not everybody wants to think, not everybody wants to do that kind of a task. There are some people who are very happy about typing away. There are some people who are very happy about various kinds of what we would call non-intellectual tasks. And that's not to degrade it. Not everybody wants to engage themselves in intellectual artificial intelligence. To me it's very difficult, as I said in the opening remarks, to separate speech understanding, communication with an intelligent entity, from the idea of using speech recognition as a tool to reduce cost effectiveness or what. There are a number of areas we could go in with this kind of stuff, and enterprising people, I suspect, hopefully will generate some ideas from this. We have time for a short question.

Q: Roland Paine, Systems Control: You identified this particular program. Would you enumerate some of the others where you are going to be doing more basic and exploratory research with speech technology as affects training in your Center?

A: We would like to explore in some way, artificial intelligence, the kinds of things that have been talked about the past two days, and we're constrained by financial reasons. In general we'd like to see these kinds of systems utilized in training.

Q: Michael Nye, Marketing Consultants: I have a question, but I wanted to make a comment concerning what Raj Reddy said, and that is that I personally believe that one of the limitations or one of the reasons why speech hasn't really, as you can say, taken off in an application environment is that too many times researchers have looked at the conceptual approach without taking a real world appreciation for economics and at such time when economics are presented that there is a cost benefit. Industry and government applications will come forth very quickly. That's a personal input although I agree with what Raj said. I just wanted to make that comment. My question is when you started in your experimentation of your system, you had some preconceived notion of what you expected, what the limitations and capabilities of this kind of system would be. I'm curious about, based on a few months of practical hands-on experience with technology that is probably limited in scope, what were the things that occurred that you did not expect that caused you to be less enthusiastic about speech understanding systems and what were the positive things that occurred that you didn't expect that made you more enthusiastic about it?

A: Some of the points were made by Mr. Herscher in his paper and I anticipate that he will make them again when he gives his presentation; they concern human factors and the man-machine interaction from a human factors standpoint, logistics of equipment, and this sort of thing. I alluded to one of those earlier about the microphone placement, and things like this. Those are the general kinds of things.

(This page intentionally left blank)

D7

N [redacted]

57-32

176343

# VOICE INTEGRATED SYSTEMS

CDR MIKE CURRAN, PH.D.[1]

NAVAL AIR DEVELOPMENT CENTER
WARMINSTER, PENNSYLVANIA

## VRAS - A Voice Recognition and Synthesis System

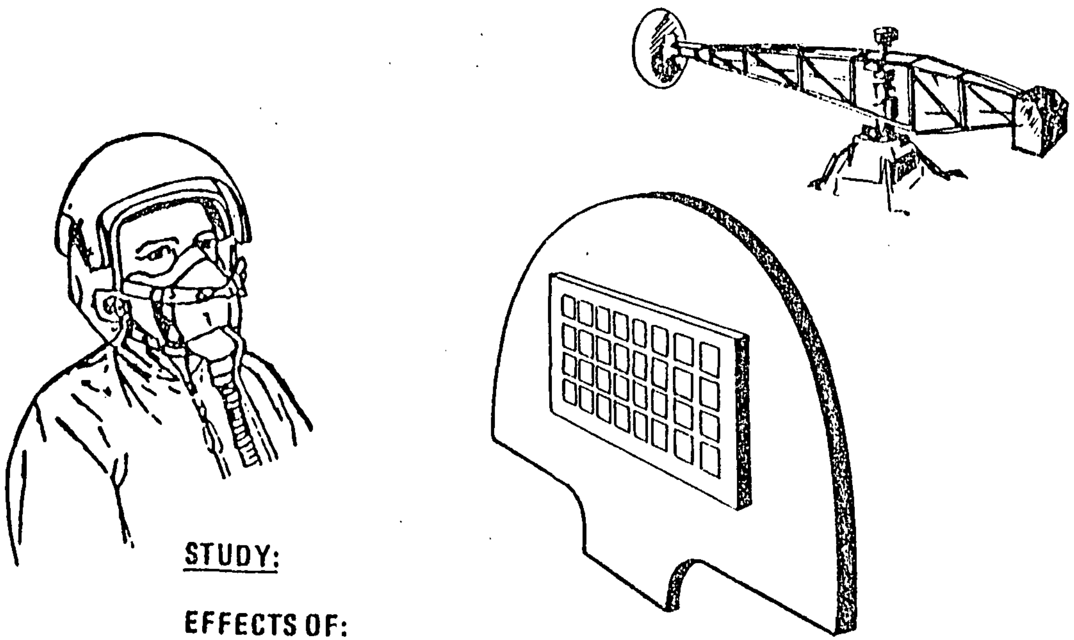
The program at NADC was initiated to determine the desirability of interactive voice systems for use in airborne weapon systems crew stations. To accomplish this effort, a voice recognition and synthesis system (VRAS) was developed and incorporated into a human centrifuge. The speech recognition aspect of VRAS was developed using a voice command system (VCS) developed by Scope Electronics. The speech synthesis capability was supplied by a Votrax, VS-5, speech synthesis unit built by Vocal Interface. The effects of simulated flight on automatic speech recognition were determined by repeated trials in the VRAS-equipped centrifuge. The relationship of vibration, G, $O_2$ mask, mission duration, and cockpit temperature and voice quality was determined. The results showed that: 1) voice quality degrades after 0.5 hours with an $O_2$ mask; 2) voice quality degrades under high (± 0.3G) vibration; and 3) voice quality degrades under high levels of G. The "voice quality" studies are summarized in Figure 1. These results were obtained with a baseline of 80 percent recognition accuracy with VCS.

The next phase of the development program called for improvement of the VCS system. This was accomplished in two ways. A consistent bit was incorporated into the process wherein reference patterns are established to improve recognition accuracy. Improved recognition accuracy was noted. A syntactical handler was developed to facilitate use of the isolated word VRAS system and to assist simultaneously in the understanding process. The developed syntactical handler was tested with teletype input and was operational with 100 percent accuracy in real time.

The major components of the VRAS system and its general operation are shown in Figure 2. We see that the spoken words, originated by a speaker, are analyzed and sent to the "Statement Understanding" component. Once the meaning of the statement is understood, it is forwarded to the "Message Handling" unit which is responsible for all exchanges of information between the VRAS system and the system computer to which it is interfaced. Having accomplished the intent of the speaker's statement, the appropriate reply is created by the "Response

[1] The opinions expressed here are those of the author and do not necessarily reflect the official policy of the United States Navy.

[redacted]



STUDY:

EFFECTS OF:

| | |
|---|---|
| COCKPIT TEMP. | COOL AVERAGE WARM |
| VIBRATION | 0 .15 .30 |
| G | 1. 2.0 4.0 |
| $0_2$ MASK | YES NO |
| MISSION DURATION | 90 MIN. |
| WORDS SAID | 1 – 28 |

FINDINGS:

VOICE QUALITY DEGRADES AFTER ½ HR. WITH $0_2$ MASK
VOICE QUALITY DEGRADES UNDER HIGH (.3 G) VIBRATION
VOICE QUALITY DEGRADES UNDER HIGHER LEVELS OF G*

* MAY BE ATTRIBUTABLE TO MASK SLIPPAGE

Figure 1. Summary of Voice Quality Studies

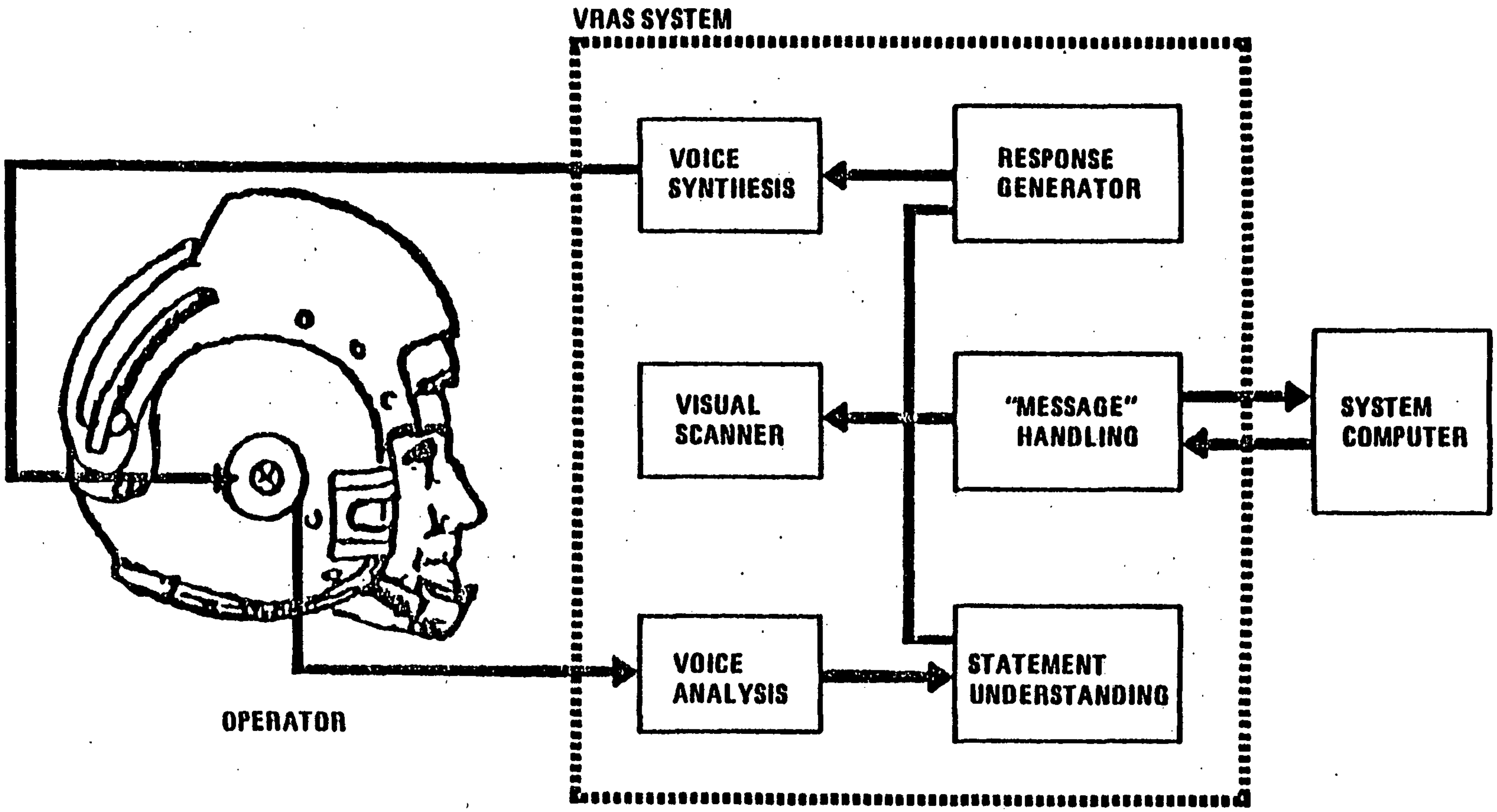


Figure 2. Voice Recognition and Synthesis (VRAS) System

Generator" and then given to the original speaker via a "Voice Synthesis" unit. The "Visual Scanner" permits visual feedback to the speaker from a variety of other VRAS units which allows the speaker to monitor visually what is being said and understood by those various VRAS components. The VRAS system also includes a printer, card reader and disc drive for logging out all communications, inputting vocabulary data, and storing different speakers' trained words for word recognition purposes.

An overview of the VRAS system, and the statement understanding approach it employs, is presented in a paper entitled "VRAS - A Voice Recognition and Synthesis System" which appears in volume VII of the 1976 SID International Symposium Proceedings. This paper was authored by Dr. Robert J. Wherry, Jr., who originated and developed the VRAS system.

This system permits the use of medium-sized vocabularies (250 words) and highly flexible statement formats. Among the unique concepts featured in the VRAS system are: 1) a "universal" statement format, 2) the use of "truth" logic to eliminate words which can no longer be appropriate in the sentence being said; and 3) a "dictionary of meaning" which permits the VRAS user to communicate a given message in a large variety of different sentences. Since the syntactical handler only requires a recognized word or phrase as input, it can be used with recognition devices other than the Scope VCS. The flexibility of the VRAS system allows for the development of a syntactical handler to accomplish any specified application within a month. The value of the syntactical handler is that it will allow the user to vary the syntactical arrangement of words during data entry without affecting recognition accuracy. Thus, the natural quality of speech as a data entry means is preserved.

The development of the VRAS facility and the VRAS concept has resulted in a powerful approach for accomplishing voice recognition and synthesis. However, since the the programming language used was at the assembly level, and since the computer employed was a Raytheon 704, only the Naval Air Development Center could readily utilize the VRAS capability. Because of the use of assembly level language, changes and improvements to VRAS have proved extremely time consuming and costly. Because of the use of the Raytheon computer the VRAS approach has not been readily applicable to the requirements of other identified voice development efforts. To correct these deficiencies a work effort has been completed which developed, tested and documented a FORTRAN IV packaging of the VRAS program. Program modifications or transferral to other computer systems or recognition and synthesis equipment have become simplified and readily implementable since all coding, except for the specific equipment interface routines, are in ANSI Fortran. Eight types of programs are required to ensure flexibility and inter-system compatibility, as well as to accomplish the VRAS syntax processing function. Seven system implementation programs are interfaced through a supervisory program that provides the few instructions required to operate the system. The programs, as shown in Table 1, can be run independently or through the supervisory program. A more detailed description of the use of each program, and how it is accessed, is provided in a report which is available for general distribution.

## "Unlimited" Vocabulary Recognition and Understanding

The thrust of the previously described VRAS program was to concentrate on understanding the meaning of what was being communicated rather than merely on the particular sequence of words which was employed. While the VRAS system does permit the use of medium-sized vocabularies, and while it does permit a relatively flexible sentence structure, the greatest single drawback to the use of real-time voice recognition and synthesis today is still the limitation on vocabulary size and sentence structure. To understand the nature of this limitation the two major approaches to word recognition or "voice analysis" must be presented. One approach involves an analog-to-digital conversion of the input voice signal and a frequency analysis using bandpass filters to record what the voice signal was during a given period of time. Correlating the obtained and expected activity levels for the different bandpass filters over time permits word recognition to occur. A second approach also uses time samples of activity for a bank of bandpass filters. In this case the patterns of activity in the filters are compared against a set of phonetic features to determine the presence or absence of various kinds of sounds (fricatives, stops, etc.). When using small, tailored vocabularies, both approaches tend to do a very good job of correctly identifying the actual word being said by the speaker.

The concept of permitting the speaker to use an "unlimited" vocabulary - any legitimate English word - has been rejected as an unrealizable near-timeframe objective for voice recognition and synthesis systems because of the difficulty in word recognition for a relatively few words. It is not merely that as the number of words in the vocabulary increases the more probable it is that two words will sound alike; it is more than this problem, which we might call the "word confusion" problem, which has discouraged the development of truly "large vocabulary" voice recognition and synthesis systems. For example, for each word in the vocabulary, its "recognition vector" (the way the speaker has previously said that word) must be stored, and if that word is to be used as a synonym for another word, then its definition must also be stored. With the VRAS system, using the Scope device, 256 bits of storage for each word in the vocabulary were required just to store the recognition vector. Assuming 16 bit computer words, each recognition vector would require 16 computer words; a vocabulary

TABLE 1

VRAS SYSTEM PROGRAMS

| | |
|---|---|
| VRAS | VRAS System Interface Program allows access to all other programs in the VRAS System. |
| TRAINING | Trains VRAS to a specific speaker and vocabulary. |
| PARSE | Processes sentences entered by speaker. |
| VOCAB | Lists the current vocabulary and subsidiary programs. |
| CONFUSIN | Enables the user to determine an appropriate recognition correlation threshold and a list of possibly confusing words. |
| RAWDATA | Enables the user to obtain listings of both the short (processed) and long (unprocessed) recognition vectors. |
| DEMO | Demonstrates the VRAS training and recognition logic. |
| RETRAIN | Allows the user to add additional repetitions of the words in the vocabulary to the composite recognition vectors stored on the disc. |

of only 2000 words would require 32,000 words just to store the recognition vectors. Another accompanying problem with large vocabularies is the increased processing time required for the additional comparisons to be made when trying to determine which word has been said by comparing the "spoken vector" with the various "recognition vector." Thus, present word recognition technology cannot handle "unlimited" vocabularies for three very good reasons: 1) large vocabularies require too much storage, 2) large vocabularies require too much processing time, and 3) large vocabularies permit too many words which tend to get confused with each other because they produce too similar recognition vectors.

While the above reasons would, at first consideration, seem sufficient to reject the concept of an unlimited vocabulary voice recognition and synthesis system, it will be seen that an alternative approach to word recognition may be possible. The alternative approach, which will be referred to as the "word-part" approach, is based on the concept that the vast majority of words used by speakers are various combinations of relatively few prefixes, stems, and suffixes. If an incoming word can be analyzed into its component word-parts, not only can the word be correctly recognized, but also its appropriate meaning can be established without reference to a "dictionary of meaning."

Just as the stem of the word has its own meaning, so also do the prefixes and suffixes. It, therefore, becomes possible to conceive of a new word recognition approach which analyzes each spoken word into its component prefixes, suffixes, and stems; to determine the meanings of these components; and to use those component meanings to determine what the speaker is saying without ever attempting to recognize the whole word or to have a definition of the whole word stored in memory.

This new approach to an "unlimited" vocabulary voice recognition and synthesis system will be pursued during fiscal years 1978 and 1979 as an exploratory development effort which should complement and extend the previously described VRAS development program.

## Integrated Applications of Automated Speech Technology

Progress in isolated word recognition, syntactical handling, and other speech technology areas provides evidence to suggest that the initiation of a Navy Advanced Development Program is justified. However, developments and progress in separate speech technology areas can only achieve their true potential if and when they are successfully integrated into total system applications. It is noteworthy that several such integrations have been achieved. The Naval Training Equipment Center's Ground Controlled Approach Controller Training System has utilized speech recognition to effect control of an aircraft/pilot simulation, and to provide the basis for the objective performance measurement of the trainee's behavior. The Department of Transportation's Automated Command Response Verification System has demonstrated the integration of automated speech technologies (AST) in an operational ship-safety role. In these and other government applications AST has done more than make life a little simpler. A number of applications have successfully demonstrated the power of AST-based systems in solving problems that could not have been addressed before the emergence of these technologies. Therefore, it seems that an advanced development program which synergistically draws upon the results of past and present AST efforts is a reasonable and worthwhile next step.

However, before such a program can be initiated and successfully pursued, several information gaps must be resolved. Specifically, if AST is to be applied to the areas of airborne crew station design, performance measurement and training simulation, several new methodologies must be developed. They include: 1) a method for identifying high payoff applications of voice interactive systems in terms of the enhancement of both operator and system performance; 2) a methodology for assessing the technical feasibility of AST for each proposed application; and 3) a methodology for integrating the above information sources and generating a rationale for mutually supportive basic research, and exploratory and advanced development requirements.

The Integrated Applications of Automated Speech Technology was an exploratory development program initiated in fiscal year 1977 to develop these methodologies. This program will be completed in early fiscal year 1978. The program objective is to develop methodologies for integrated applications of automated speech technologies for Navy system development, training, and operational settings. The program approach includes five major tasks: 1) review government applications of AST; 2) perform crew station design analyses; 3) examine performance measurement capabilities; 4) examine training applications; and 5) prepare a program/implementation plan.

The objective of the review task was to review critically the present applications of AST, and their supporting data, to establish a baseline of present progress from which the Navy can draw to plan AST applications. The completed review identified present capabilities and advancements, as reflected in successful system applications of AST, for type of speech recognition (i.e., isolated and limited continuous), vocabulary, recognition accuracy, syntactical handling, and user acceptance.

The crew station design, performance measurement, and training applications tasks have addressed documentation available for the Navy P-3C anti-submarine aircraft weapon system to develop the desired methodologies. For the crew station design task the "on station" mission segment was examined for each crew member by considering tasks to be performed and subsystems to be exercised. After consideration of the variables that affect the application of voice technology to crew station design, a four dimensional rating system was developed. The dimensions included: the technical feasibility of implementing voice for accomplishing the task; the utility of implementing voice to accomplish the task; time/accuracy requirements for the task; and the impact of unassessed variables such as aircraft noise and mission duration. Using this rating strucutre, each task was reviewed utilizing the four dimensional requirements, and assigned a four digit code of numbers corresponding to the four task requirement. For each task the four digit code was reduced to a one digit code corresponding to initial AST payoff. Previously obtained criticality and frequency ratings for each task were applied to this initial AST payoff rating to obtain an overall AST payoff rating. Finally the ratings for the tasks were converted to a matrix format. As an example of this process, Table 2 shows tasks by subsystems for the P-3C Pilot. The AST ratings for all detailed tasks to be included within a matrix cell were treated statistically to determine a single AST potential payoff rating for each matrix cell. Table 3 summarizes the most promising voice applications areas for both the P-3C Pilot and TACCO. It should be recalled that the objective of this task was not to identify voice applications for the P-3C, but to develop the methodologies required to identify high payoff applications of voice technology.

As of this time the performance measurement and training applications tasks are not completed.

The last task of this effort involves preparation of a program/implementation plan. The general approach for integrating various information sources and generating a rational for research implementation requirements is presented in Figure 3. Neither the various trade-off analyses nor the methodology for effecting the integration of the various information elements have been completed. Again the promise of this task is that the approaches developed for the generation of the trade-off analyses and the integration of the information sources will be incorporated when attempting to apply voice technology to other airborne systems/subsystems. The identification of technology voids and problems will serve as the basis of an interlocking technology base program covering the full spectrum of basic research through advanced development.

## VIST - Voice Interactive Systems Technology

VIST is a new advanced development program being initiated in fiscal year 1978. It is viewed as the application or implementation of the products obtained from the previously described AST exploratory development effort. The objective of the program is to demonstrate the

TABLE 2

VOICE TECHNOLOGY PAYOFF FOR P-3C PILOT

POSITION: P-3C PILOT

MISSION SEGMENT: ON STATION

POTENTIAL PAYOFF CODE:

| | | |
|---|---|---|
| 1. | GREEN | HIGH |
| 2. | BLUE | MEDIUM |
| 3. | YELLOW | LOW |

SUBSYSTEMS

| GENERIC TASKS | PROPULTION | FUEL | ELECTRICAL | HYDRAULIC | ECS | AIR FRAME | FLT CONTROLS | FLT INSTRUMENTS | COMMUNICATIONS | NAVIGATION | ECM | ASW: | SEARCH STORES | ON-TOP POS. IND. | BT & SEA NOISE | MAD | LLLTV | PHOTOGRAPHIC | CASS | DATA HANDLING & DISPLAY | SEARCH RADAR | ARMAMENT | CREW |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MONITOR INDICATOR | 1 | 2 | | | | | | | 1 | | | | | | | | | 2 | | | | 2 | |
| MONITOR SITUATION | | | | | | | | | | | | | | | | | | | | | | | |
| PERCEIVE DATA | | | | | | | | | | | | | | | | | | | | | | | |
| FORMULATE PLAN | | | | | | | | | | | | | | | | | | | | | | | |
| DETERMINE SOLUTION | | | | | | | | | | | | | | | | | | | | | | | 1 |
| ENTER DATA | | | | | | | | | | | | | | | | | | | | | | | |
| RECEIVE DATA | | | | | | | | | 3 | | | | | | | | | | | 3 | | | |
| COORDINATE DATA | | | | | | | | | | | | | | | | | | | | | | | |
| ACTIVATE CONTROLS | 3 | | | | | | | | 2 | 3 | | | 2 | | | | | 2 | | 3 | | 2 | |
| ADJUST CONTROLS | | | | | | | | | 2 | | | | | | | | | | | 3 | | 3 | |
| PERFORM MANEUVER | | | | | | | | | | | | | | | | | | | | | | | |

## TABLE 3

## P-3C PILOT/TACCO SUMMARY RESULTS

### PRELIMINARY ANALYSIS OF MOST PROMISING AREAS FOR VOICE APPLICATION

A. TASKS (IN ORDER OF PROMISE)

1. MONITOR INDICATORS (ALERTS)
2. ACTIVATE SWITCHES (FUNCTION SWITCHES)
3. ENTER DATA (KEYBOARD)
4. ADJUST CONTROLS (KNOBS)
5. RECEIVE DATA (CRT TABLEAUS)
6. COORDINATE DATA (COMMUNICATION)

B. SUBSYSTEMS (IN ORDER OF PROMISE)

1. COMMUNICATIONS
2. PROPULSION
3. SEARCH STORES
4. PHOTO
5. DATA HANDLING
6. ARMAMENT
7. CREW

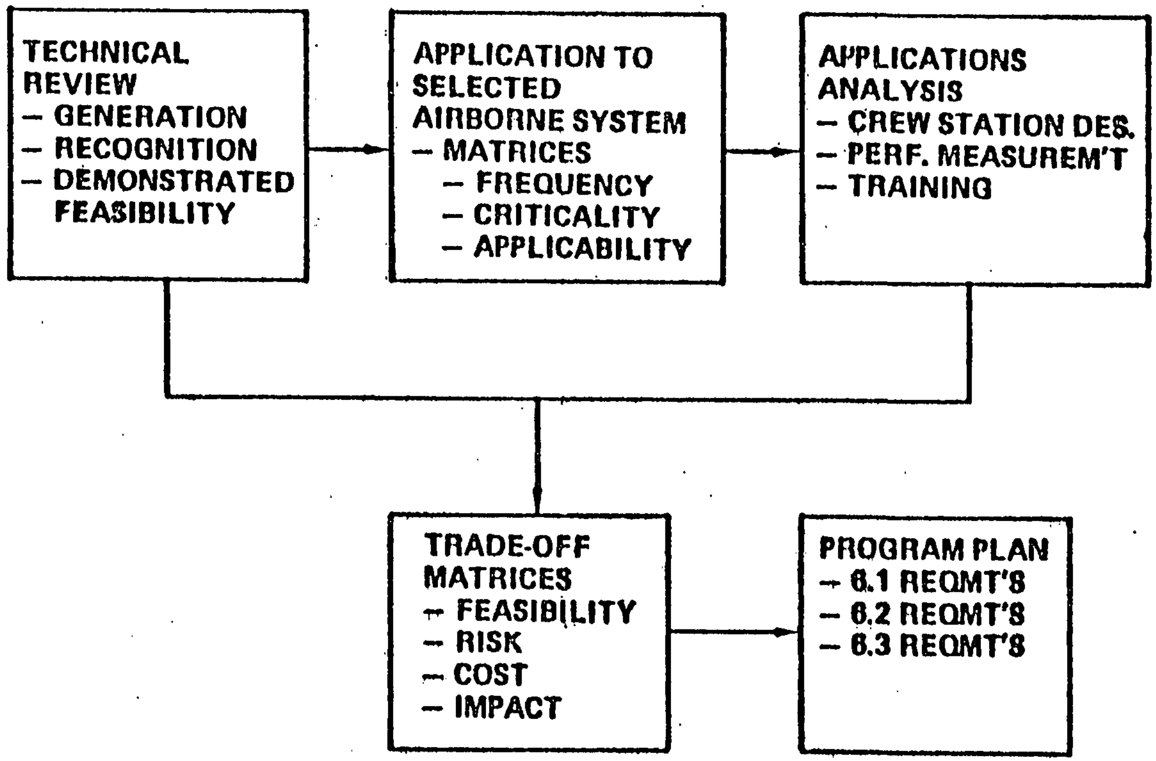


Figure 3. AST Program Plan Rationale

applicability of voice-based technologies to the specific areas of airborne controls/displays design, performance measurement, and training simulation. The general purpose of the effort is to apply AST to obtain a carefully determined voice interactive system (VIS) which, when coupled with a multi-task simulator will: 1) reduce operator loading by sharing operator display/control functions between the visual/motor channels and the verbal/auditory channels, and 2) allow direct measurement of operator performance which will result in more precise control of the training and skill monitoring processes, and a more meaningful index of operational readiness.

The general approach of this program involves: 1) incorporation of the results of the exploratory development Integrated Applications of Automated Speech Technologies effort to develop a strategy for demonstrating the suitability of applying voice to accomplish program objectives; 2) design, development, and exercising of a voice system and simulator to allow for performance of selected airborne tasks, and the provision of a capability for recording and processing operator/ system voice transactions; 3) exercising the voice system and simulator to evaluate proposed task applications for cost effectiveness, contribution to system effectiveness, and operational acceptability; and 4) generation of detailed system specifications for the implementation of voice applications by program managers.

A five year development effort is spelled out in the detailed program plan. Major milestones include: 1) determination of voice system and simulator functional requirements; 2) determination of a configuration for the demonstration system; 3) preparation of the detailed work breakdown structure; 4) preparation of the detailed implementation plan; 5) development and integration of the demonstration system; 6) demonstration/evaluation of the system to determine the adequacy of each implemented task to fulfill functional requirements; 7) performance of required cost/benefit/effectiveness analyses; 8) generation of detailed design specifications for each major component of the voice system, e.g., a detailed design specification for a voice recognition element; 9) generation of a taxonomy of selected airborne tasks which can serve as a guide for the utilization and application of voice technology; and 10) transition to engineering development. This last effort calls for the identification of target or candidate systems, and system tasks, which promise high payoff for the application of voice technology. The advanced development program will also include integration of the prototype voice system with selected platforms to provide intermediate demonstrations of the utility of voice technology for target or candidate systems, and to provide for an orderly transition to engineering development.

The generalized product of this effort is an intermediately-validated voice interactive system which has demonstrated its capability for providing a solution for problems of airborne control/display design, performance measurement, and training simulation. Upon completion of the VIST program a data base consisting of: 1) a set of detailed design specifications, and 2) a set of identified airborne tasks will be available to program managers responsible for the development of major airborne systems/subsystems. Such managers can match the tasks to be performed by the operators of their system or subsystem against those tasks which have been shown to contribute to enhanced system effectiveness by the application of voice technology. More important, program managers will have available a vehicle to accomplish the implementation of voice technology for their particular purpose. That is, they will have available detailed design specifications for voice interactive systems which, if identified as system requirements, will provide sound assurance that the developed system will best reflect the maximal contribution of voice technology.

In addition, the voice integrated system and simulator capability developed during the advanced development program can serve as an in-house capability to be exercised to support a program manager in determining whether a contractor-developed voice system adheres to system design and performance requirements. A program manager will have available for his use a tool to accomplish independent system verification and validation. This approach has repeatedly demonstrated its value in the development of large-scale software and hardware systems.

While the VIST program is directed toward implementation in airborne systems, the processes and products of this program should be readily utilizable for accomplishing the incorporation of voice technology in complex operator stations in submarines, surface ships, and ground-based installations which require operator and system interfacing.

## BIOGRAPHICAL SKETCH

### CDR Mike Curran, Ph.D.

Commander Mike Curran is Block Program Manager for human factors engineering exploratory development, and Head, Technology Development Branch, at the Naval Air Development Center (NADC), Warminster, PA. He received his B.A. from U.C.L.A. in 1961, his M.A. in 1964 from George Peabody College, and his Ph.D. from Texas Tech University in 1974. In addition to his involvement in Navy technology development Commander Curran has served as coordinator of all crew systems support for the LAMPS MK III Weapon System at NADC. While at the Pacific Missile Test Center, Pt. Mugu, CA., he was Head, Human Factors Engineering Branch (1970-1972), and also conducted both exploratory development simulation studies of new airborne display concepts and supported the human factors engineering test and evaluation of the A-7E and F-14A Weapon Systems (1968-1972). From 1963-1968 Commander Curran was stationed at the Naval Aerospace Medical

Research Laboratory, Pensacola, FL., where he contributed to the development of an on-line, interactive test system to be employed as a research tool and as an approch for implementing secondary selection procedures. He also pursued a long-term program concerned with the effects of psychological stress on performance.

## DISCUSSION

## CDR Mike Curran, Ph.D.

Q: Bob Hilgendorf, Reconnaissance Strike System Program Office at Wright Patterson: You mentioned that your FY78 VIST Program is a 6.3 effort. Was your FY77 program where you were developing some of these methodologies a 6.3 effort also?

A: No, maybe it was confusing to you. The Integrated Applications of Speech Technology, i.e. the five tasks I described, will flow into our advanced development program.

Q: Hilgendorf: Okay, so it's a 6.2 effort?

A: Yes.

Q: Hilgendorf: Who is doing the work?

A: Who is doing the integrated applications work? Logicon and Boeing on a teamed contract.

Q: Hilgendorf: Okay, you were talking about the generation of some systems specifications as a fallout to your 6.3 program.

A: No, as the product, not the fallout.

Q: Hilgendorf: Okay, as a product. Do you have any time factors as to when you have pitched to your people that you are going to be producing these specification?

A: I would love to give that one to Cdr. Lane. Depending on the availability of funding, what is the current program schedule, Cdr. Lane?

Cdr. Lane: Fiscal Year 81 or 82.

Q: Hilgendorf: Okay, then my last question is even though the decisions concerning the applications to certain airborne platforms are supposed to be made in an orderly fashion, do you have any teasers as to what systems you're talking about besides perhaps P-3C?

A: My own private opinion?

Q: Hilgendorf: Yes, that's good enough.

A: V/STOL-C. And only if you know that there is a V/STOL-A and a V/STOL-B. It's a pretty safe bet. We can make an impact on V/STOL-C, if there continues to be a Navy V/STOL Program. That time frame is approximately 1990.

Q: Hilgendorf: Okay, I don't want to comment any further.

Q: John Allen, MIT: I wonder in the application you mentioned where you talked about computing the meanings of words from route words and prefixes and suffixes, what kind of vocabulary you have in mind and do you have a basis for determining those meanings from the constituent morphemes of the word?

A: Yes. I'll make one statement. Since this effort is in the processing of contracting now, I can't go into any more detail. We are not going to be dealing with the total unlimited English language dictionary of 80,000 words. Our effort will restrict itself to what we call the aviation dictionary. We think aviators do not utilize more than 10,000 words, whether they're Navy or Air Force. We have already gone though a College Dictionary and know there is a large number of words that are never used in aviation. We're talking, initially, of an unlimited vocabulary of 10,000 words. Those words, of course, include prefixes and suffixes. You may be talking about 1,000 to 2,000 parts, or what we'd call word roots.

Q: Ken Woodruff, Systems Research Laboratories: Cdr. Curran, you talked about the Boeing effort...

A: Boeing - Logicon effort.

Q: Woodruff:....tried to develop a methodology for determing high payoff applications and I believe they are using the P-3C as their test bid for that. You did not get into the criteria by which they are making those payoff decisions. Would you care to comment on that? The reason I ask the question is the P-3C is a multi-crew vehicle. We have a severe problem within the Air Force community that we are going to all single seat aircraft. The priority structure might change considerably if that were your consideration.

A: Even though Boeing and Logicon are here, I think I can comment on that since our first decision in initiating this effort was what platform we were going to use. Obviously, in the Navy we have other than one-seat airplanes. But our decision regarding a candidate platform to develop these methodologies involved do we use fighter, do we use attack, do we use multi-engine? By the way, Commander Hanson is here. This effort is jointly sponsored with the Office of Naval Research. There was considerable debate as to what type of platform we should examine. Should it be ASW, fighter, should we use helo or what? I think by browbeating or consensus, we decided on ASW because we in the Navy at NADC think we know something about ASW. Boeing has a wealth of experience in P-3's and Logicon is close to the Navy environment and quite expert in voice systems applications. So we chose something we are familiar with, but also something that is well-documented. The P-C Charley is a well-decumented aircraft in terms of what the tasks are. The other approach was we could have chosen a future platform, and tried to anticipate tasks. It would have been more difficult. So, that's why we chose the P-C Charley looking at all operators and all tasks.

Q: Unidentified Questioner (in distance): Question not recorded.

A: You are going to get into the complex meaning of the technology payoff, the utility. I brought copies of the rating process which describes in detail how we came up with the final "AST Payoff Rating." And I'll be happy to share it with you. Any other questions?

Again, just let me emphasize just one thing. You saw final one-digit ratings and heard my verbal description which was cursory. You can realize that with the P-C Charley we had available a Coarseware documentation on criticality and frequency for operator tasks. That information, combined with Boeing and Logicon's best judgment of what the technical feasibility of the application was, gave us the worth of the proposed. Any other questions?

[redacted]

# AUTOMATIC SPEECH RECOGNITION RESEARCH AT NASA-AMES RESEARCH CENTER

Clayton R. Coler
NASA-Ames Research Center
Moffet Field, CA 94035

Robert P. Plummer
University of Utah
Salt Lake City, UT 84112

Edward M. Huff
NASA-Ames Research Center
Moffet Field, CA 94035

Myron H. Hitchcock
Computer Sciences Corporation
Mountain View, CA 94043

Automatic speech recognition (ASR) is being investigated at NASA-Ames Research Center as part of a broad program in Flight Management Systems research. The goal of the Flight Management program is to develop a base of practical knowledge and experience concerning pilot information and control requirements for future highly automated commercial flight systems [1]. The motivation for this research is concern that the air traffic environment is becoming highly congested and will eventually become saturated unless some means is found to improve the overall precision and scheduling of flights, particularly in the dense terminal area.

Various display and control devices are being investigated as aids to the crew of future aircraft. Potential information displays include multifunction, area-navigation, moving map, collision warning, traffic situation, ground proximity, system status, and various attitude displays. Virtually all of these displays involve <u>selection features</u> such as orientation, scaling, symbology, or numerical parameter options that may be left to the pilot to determine. Option selection will be determined in part by the ease or difficulty that the pilot will have in making his choices, and the degree to which these activities may interfere with more important pilot functions. Speech technology offers opportunities for increasing the effectiveness of pilot-system interaction, and both speech recognition and speech synthesis are being considered: the former as a potential input alternative to numerous and complex keyboard arrangements, and the latter as an optional output medium for presenting flight-critical information. This paper describes speech recognition work at NASA-Ames Research Center. Near-term ASR testing and evaluation in a motion simulator, and subsequent

Page Intentionally Left Blank

(ties are resolved arbitrarily). Assuming that the vocabulary words are equally likely to occur and that all misclassifications are equally costly, the maximum likelihood decision is obtained by using as the discriminant functions

$$g_i(X) = p(X|i), \quad i = 1, \ldots\ldots, N.$$

That is, the ith discriminant function is the likelihood of pattern, X, with respect to category, i. Since the patterns are 120-bit binary vectors, and assuming the pattern components to be statistically independent, then

$$g_i(X) = \prod_{j=1}^{120} p(x_j|i), \; i = 1, \ldots, N, \text{ where}$$

$X = (x_1, \ldots\ldots, x_{120})$ and each $x_j$ is 0 or 1.

For computational purposes, the discriminant functions actually used are

$$g_i(X) = \log \prod_{j=1}^{120} p(x_j|i) = \sum_{j=1}^{120} \log p(x_j|i).$$

The logarithms, which are computer by table look-up, do not affect the classification since the log function is monotonic-increasing.

It remains to state how the probabilities, $p(x_j|i)$ are found. This is done in advanced by collecting training samples, $Y^{i,1}, \ldots Y^{i,m}$ for each vocabulary word, i, and using the following as estimates for the probabilities:

$$p(x_j = 1|i) = \frac{\sum_{k=1}^{M} y_j^{i,k}}{M}$$

$$\text{and } p(x_j = 0|i) = 1 - p(x_j = 1|i).$$

The number of training samples, M, is usually between 10 and 25.

Three additional features have been added to the maximum likelihood algorithm. The first derives from the fact that even practiced speakers change some of their pronunciations slightly over time. Compensation from these changes may be accomplished by updating the discriminant function probabilities. In situations where the algorithm receives feedback stating the correctness of its classifications, then if pattern, Z, is correctly classified as belonging to category, m, the probabilities $p(x = 1|m)$ are replaced by

$$a \cdot p(x_j = 1|m) + (1-a) \cdot z_j \;, \text{ for } 0<a<1 \text{ and } j=1, \ldots, 120.$$

This exponential smoothing technique can be used to make the algorithm more or less responsive to changes in pronunciation by varying the value of the variable a. A typical value is 0.94.

A second feature allows the algorithm to "reject" utterances for which classification is uncertain, rather than risk misclassification. The measure of uncertainty is a simple one: the ratio of the second highest score to the highest. The nearer the ratio to 1.0, the more uncertain the classification. The threshold at which rejection takes place is a parameter of the algorithm.

Finally, in many flight systems applications, the command language spoken by the user of the speech recognition system has a simple (finite-state) syntax. For example, after the command "landing gear," the only meaningful utterances might be "up," "down," and "status." By doing recognition only against the subset of vocabulary words that is valid at each point in a command string, both the accuracy and efficiency of the algorithm are increased. A tree-like structure is used to associate with each vocabulary word that set of words that can follow it in a command string. As each word is recognized, the result is used to guide the traversal of the tree.

The current version of the Ames speech recognition system runs in real time on a PDP 11/10 computer. Encoding an utterance into 120-bit form requires about 0.3 seconds, and recognition requires an additional 0.2 seconds for small vocabularies or subsets.

## RECOGNITION ALGORITHM EVALUATION

A standard set of data was needed for evaluation of the various recognition algorithms. To meet this requirement, 20 untrained male speakers used the 10-word digit vocabulary to provide a total of 1250 utterances for each speaker. The data for each speaker were collected in 5 blocks of 250 utterances each during a single 1 hour 30 minute session.

During data collection, the 120-bit pattern derived from each utterance was transmitted directly from the VCS to a PDP-12 computer for storage on magnetic tape. All data were later transferred to a Xerox Sigma 9 computer where the various recognition algorithms were used to process the data. During data processing, the first block of data for each speaker was used to train the recognition software and then recognition was attempted on the remaining 4 blocks of data for that speaker; thus 1,000 recognition utterances were processed for each speaker. The same data were repeatedly processed to provide independent evaluation of each recognition algorithm. The results are shown in Table I.

Pg. 144 ✓

TABLE 1

## SPEECH RECOGNITION ACCURACY FOR UNTRAINED GROUP: 10-WORD (DIGIT) VOCABULARY

(20 SPEAKERS; 1000 UTTERANCES EACH SPEAKER)

| RECOGNITION ALGORITHM | PERCENTAGE CORRECT | | |
|---|---|---|---|
| | MEAN | STANDARD DEVIATION | RANGE |
| $VCS_{(5)}$ | 87.6 | 6.4 | 64.5 – 94.4 |
| $SW_{(5)}$ | 95.0 | 2.0 | 91.1 – 98.6 |
| $SW_{(10)}$ | 96.9 | 1.7 | 93.4 – 98.9 |
| $SUBMAX_{(10)}$ | 99.4 | 0.3 | 98.8 – 99.8 |
| $SUBMAX_{(25)}$ | 99.5 | 0.3 | 98.7 – 99.9 |
| $SUBMAX_{(25)}$ REJ* | 99.9 | 0.1 | 99.6 – 100.0 |
| *PERCENTAGE REJECTED: | 5.0 | 2.3 | 1.7 – 9.4 |

The recognition algorithms in Table 1 are listed in order of increasing efficiency. The first algorithm, VCS(5) requires 5 training samples of each command, and simulates the hard-wired algorithm of the VCS (for comparison with initial VCS test results). This least effective recognition algorithm correctly classified 87.6% of 20,000 utterances. In contrast, 99.9% of 19,000 utterances (1000 utterances were rejected) were correctly classified by the most effective algorithm, designated SUBMAX(25)REJ (incorporating all the features discussed above and trained on 25 samples of each word). Thus a mean improvement of 12.3% correct was gained over the basic recognition technique. Syntax was not involved in any of the 10-word testing; all recognition was done against the entire vocabulary.

The greatest sequential improvement, 7.4% correct, was gained by use of the algorithm designated SW(5), a maximum likelihood algorithm without rejection or updating, and trained on 5 samples of each word. Doubling the size of the training data set with the same algorithm, SW(10), produced an additional gain of 1.9% correct. Use of rejection and updating with the same size training data set, algorithm SUBMAX(10), provided a further gain of 2.5% correct, and use of all available training data, algorithm SUBMAX(25), provided an additional gain of 0.1% correct. By using algorithm SUBMAX(25)REJ, which rejected the 5% of utterances with the greatest uncertainty of classification, an additional 0.4% correct was gained. (The value of the rejection threshold that produced 5% rejection was determined empirically.)

Both the standard deviation and the range of mean percentage correct (speaker with highest mean minus speaker with lowest mean) consistently decreased as increasingly efficient algorithms were used. The standard deviation decreased from a maximum of 6.4% to a minimum of 0.1% correct, and the range decreased from a maximum of 29.9% to a minimum of 0.4% correct. Large relative reductions in standard deviation (from 6.4% to 2.0% correct) and range (29.9% to 7.5% correct) resulted from use of the SW95) algorithm. The largest relative reductions in standard deviation (from 1.7% to 0.3% correct) and range (from 5.5% to 1.0% correct) resulted from use of the SUBMAX(10) algorithm (with dynamic updating).

The principal achievement of the Ames recognition algorithm development work is that recognition accuracy for the least successful of the 20 speakers was improved to 98.7% correct without rejecting any of his utterances. For successful flight applications, a recognition system must perform well for even the least proficient speaker.

Since asking a pilot to repeat a command seems preferable to misclassification [3], the ability to reject in cases of uncertainty is desirable. With 5% rejection, recognition accuracy was further improved to 99.6% correct for the least proficient speaker.

After achieving a high level of recognition accuracy for the 10-word digit vocabulary, it was desirable to evaluate recognition accuracy of the system with a larger vocabulary suitable for flight system use.

## RECOGNITION ALGORITHM ACCURACY ON A 100-COMMAND VOCABULARY

Although a vocabulary larger than 10 commands would be required for most flight systems applications, only a few of the commands may be needed at any given time during a mission. The commands may be assigned to different vocabulary subsets according to their sequential use in the mission. The pilot would have access to any command at all times simply by executing the proper access sequence for that command. For example, a 58-command vocabulary selected for use in a fixed-base flight simulator mission consists of 17 different vocabulary subsets; the smallest subset contains 3 commands, and the largest contains 12. A syntax structure was imposed to develop branching chains of command by sequentially combining appropriate commands from various subsets to yield a total of more than 46,000 unique and potentially meaningful sequences (from the original 58-command vocabulary). In addition to increasing the number of unique command possibilities, the syntax structuring method also reduces the number of active commands in the recognition set (min. = 3; max. = 12) at any given point in time, thus reducing the complexity of the recognition problem and increasing the probability of maintaining a high level of recognition accuracy over the entire mission simulation [4].

To evaluate recognition accuracy on a large vocabulary, a 100-command flight vocabulary constructed for use in a full mission (take-off-to-landing) simulation was selected. The 100-command vocabulary is shown in Table 2.

A group of ten untrained male speakers each used the 100-command vocabulary to provide 25 training utterances (used to train the recognition algorithms) and 100 recognition utterances of each command, a group total of 100,000 recognition utterances.

### Overall Processing Results

Recognition results for overall processing without the use of syntax are shown in Table 3. Recognition accuracy for the entire vocabulary was 93.2% correct without rejection and 95.7% correct when 5% of the utterances were rejected. Results for the 10 digits within the overall processing are shown for comparison with previous 10-digit vocabulary results.

The command language syntax for this vocabulary groups the commands into 15 subsets ranging in size from 3 to 10 commands as shown in Table 4.

TABLE 2

## 100-COMMAND FLIGHT SIMULATION VOCABULARY

1. AIRPORTS
2. ARRIVAL TIME
3. AIRSPEED
4. ALTITUDE
5. ALPHA
6. AUTO
7. AUTOPILOT
8. BETA
9. BLANK
10. BY
11. CAPTURE
12. CENTER
13. CHANGE
14. CHANNEL
15. CHART
16. CHECK LIST
17. CLEAR
18. CLIMB
19. COMMUNICATION
20. COUPLE
21. COURSE
22. DELTA
23. DEGREES
24. DESCEND
25. DISPLAY
26. DOWN
27. EAST
28. EIGHT
29. EMERGENCY
30. ENGAGE
31. ENTER
32. ERROR
33. EXECUTE
34. FIVE
35. FLIGHT PLAN
36. FLY TO
37. FOUR
38. FREQUENCY
39. GAMMA
40. GLIDESLOPE
41. GO
42. GO AROUND
43. HEADING
44. HOLD
45. HORIZONTAL
46. HUNDRED
47. I.D.
48. IN
49. INDEX
50. INITIALIZE
51. INPUT
52. LARGER
53. LEFT
54. MAP
55. MASTER
56 M.L.S.
57. MERGE
58. MINUS
59. NAV AUTO
60. NAVIGATION
61. NEGATIVE
62. NINE
63. NORTH
64. OFF
65. ON
66. ONE
67. OUT
68. OVER
69. PLUS
70. POINT
71. POSITIVE
72 PREDICTOR
73. REFERENCE
74. RIGHT
75. SCALE
76. SECTOR
77. SELECT
78. SET
79. SEVEN
80. SIX
81. SMALLER
82. SOUTH
83. SPEED
84. TERRAIN
85. THOUSAND
86. THREE
87. TIME
88. TRACK
89. TURN
90. TWO
91. UNDER
92. UP
93. VECTOR
94. VELOCITY
95. VERIFY
96. VERTICAL
97. V.O.R.
98. WAYPOINT
99. WEST
100. ZERO

TABLE 3

## SPEECH RECOGNITION ACCURACY FOR 10 UNTRAINED SPEAKERS: 100-WORD FLIGHT SIMULATION VOCABULARY

## RECOGNITION RESULTS FOR OVERALL PROCESSING

| RECOGNITION ALGORITHM | PERCENTAGE CORRECT FOR ENTIRE VOCABULARY | | PERCENTAGE CORRECT FOR DIGIT SUBSET | |
|---|---|---|---|---|
| | MEAN | RANGE | MEAN | RANGE |
| SUBMAX$_{(25)}$ | 93.2 | 89.3 – 96.8 | 91.8 | 88.4 – 93.7 |
| SUBMAX$_{(25)}$ REJ* | 95.7 | 93.2 – 98.1 | 94.7 | 92.0 – 96.4 |
| *FREQUENCY OF REJECTION | 5.0 | 2.8 – 7.1 | 6.0 | 3.2 – 8.2 |

TABLE 4

## FREQUENCY OF SUBSET SIZES:
## 100-COMMAND FLIGHT SIMULATION VOCABULARY

| SUBSET SIZE | FREQUENCY |
|---|---|
| 3 WORDS | 2 |
| 4 WORDS | 3 |
| 5 WORDS | 1 |
| 6 WORDS | 3 |
| 7 WORDS | — |
| 8 WORDS | — |
| 9 WORDS | 1 |
| 10 WORDS | 5 |
|  | TOTAL = 15 SUBSETS |

## Subset Processing Results

Recognition results for subset processing are shown in Table 5. For the entire vocabulary, subset recognition accuracy without rejection is 98.6% correct, and is higher than the corresponsing overall processing result by 5.4% correct, while the range has decreased from 7.5% to 1.8% correct. With 5% rejection, subset recognition accuracy is increased by 1.0% correct to 99.6% correct, and the range is decreased to 0.5% correct. Results for the 10 digits within the subset processing are shown for comparison with previous 10-digit results.

Tables 6, 7, and 8 show the commands, subset processing results without rejection, and rank order of recognition accuracy for each of the 15 subsets. The subset processing results shown in Tables 6, 7, and 8 are summarized in Table 9, together with corresponding results from the overall processing.

During overall processing, recognition was done against the entire 100-command vocabulary. (During subset processing, the recognition decision for each utterance was based on comparison to only 3, 4, 5, 6, 9, or 10 commands, depending upon the size of the active subset.) The overall processing results shown in Table 9 were then obtained by simply combining and averaging individual command results by subset group. (The subset groups are shown in Tables 6, 7, and 8.)

The rank order results for subset processing shown in Table 9 are in general agreement with the expectation that recognition accuracy will decrease as subset size is increased. When recognition accuracy for a subset of small size was considerably below the levels obtained with larger subsets, the disparity was usually attributable to frequent misclassification between two commands in the small subset. For example, although Subset L contains only 4 commands, the resulting mean percentage correct by subsetting was lower than corresponding results for eight subsets of larger size. The primary reason for error within Subset L was misclassification between the commands "alpha" and "delta" (see Table 7).

## Digit Subset Results

The 10-digit subset, subset D, ranged 14th of the 15 subsets in recognition accuracy by subset processing and 13th by overall processing. Of the five 10-command subsets, subset D recognition accuracy ranked 4th by both processing methods. These results demonstrate that the 10 digits comprise a relatively difficult 10-word ASR vocabulary. Since accurate recognition of the 10 digits is essential for most potential flight systems applications, the digits appear to be an excellent small vocabulary for use in recognition accuracy evaluation of an ASR system proposed for flight systems use.

TABLE 5

## SPEECH RECOGNITION ACCURACY FOR 10 UNTRAINED SPEAKERS: 100-WORD FLIGHT SIMULATION VOCABULARY

## RECOGNITION RESULTS FOR SUBSET PROCESSING

| RECOGNITION ALGORITHM | PERCENTAGE CORRECT FOR SUBSETS: ENTIRE VOCABULARY | | PERCENTAGE CORRECT FOR DIGIT SUBSET | |
|---|---|---|---|---|
| | MEAN | RANGE | MEAN | RANGE |
| SUBMAX$_{(25)}$ | 98.6 | 97.5 – 99.3 | 98.2 | 97.1 – 99.1 |
| SUBMAX$_{(25)}$ REJ* | 99.6 | 99.3 – 99.8 | 99.6 | 98.9 – 99.9 |
| *FREQUENCY OF REJECTION | 5.0 | 2.7 – 7.3 | 7.4 | 4.1 – 14.0 |

TABLE 6

## SUBSETS FOR 100-COMMAND FLIGHT SIMULATION VOCABULARY

MASTER PAGE (LEVEL 1)
**SUBSET A (6 COMMANDS)**

7. AUTOPILOT
15. CHART
16. CHECK LIST
19. COMMUNICATION
35. FLIGHT PLAN
60. NAVIGATION

98.7% CORRECT
RANK = 9

AUTOPILOT COMMANDS (L2)
**SUBSET B (4)**

4. ALTITUDE
43. HEADING
83. SPEED
96. VERTICAL

99.15% CORRECT
RANK = 3.5

NAVIGATION COMMANDS (L2)
**SUBSET C (4)**

2. ARRIVAL TIME
56. M.L.S.
59. NAV AUTO
97. V.O.R.

98.725% CORRECT
RANK = 7

COMMUNICATION AND
AUTOPILOT COMMANDS (L2)
**SUBSET D (10)**

28. EIGHT
34. FIVE
37. FOUR
62. NINE
66. ONE
79. SEVEN
80. SIX
86. THREE
90. TWO
100. ZERO

98.21% CORRECT
RANK = 14

CHART COMMANDS (L2)
**SUBSET E (6)**

1. AIRPORTS
54. MAP
72. PREDICTOR
75. SCALE
84. TERRAIN
94. VELOCITY

99.03% CORRECT
RANK = 5

TABLE 7

## SUBSETS FOR 100-COMMAND FLIGHT SIMULATION VOCABULARY

AUTOPILOT: ALTITUDE, HEADING, AND SPEED COMMANDS (L3)
**SUBSET F (9 COMMANDS)**

44. HOLD
46. HUNDRED
51. INPUT
58. MINUS
69. PLUS
70. POINT
73. REFERENCE
77. SELECT
85. THOUSAND

98.48% CORRECT
RANK = 10

AUTOPILOT: VERTICAL COMMANDS (L3)
**SUBSET G (6)**

18. CLIMB
24. DESCEND
45. HORIZONTAL
48. IN
67. OUT
89. TURN

99.15% CORRECT
RANK = 3.5

NAVIGATION: M.L.S. COMMANDS (L3)
**SUBSET H (5)**

14. CHANNEL
30. ENGAGE
40. GLIDESLOPE
64. OFF
65. ON

98.86% CORRECT
RANK = 6

NAVIGATION: V.O.R. COMMANDS (L3)
**SUBSET I (3)**

21. COURSE
38. FREQUENCY
47. I.D.

99.53% CORRECT
RANK = 2

CHART: SCALE COMMANDS (L3)
**SUBSET J (3)**

52. LARGER
55. MASTER
81. SMALLER

99.7% CORRECT
RANK = 1

CHART: MAP COMMANDS (L3)
**SUBSET K (10)**

12. CENTER
26. DOWN
27. EAST
49. INDEX
53. LEFT
63. NORTH
74. RIGHT
82. SOUTH
92. UP
99. WEST

97.98% CORRECT
RANK = 15

COMMUNICATION COMMANDS (OPTIONAL-L2)
**SUBSET L (4)**

5. ALPHA
8. BETA
22. DELTA
39. GAMMA

98.4% CORRECT
RANK = 13

TABLE 8

## SUBSETS FOR 100-COMMAND FLIGHT SIMULATION VOCABULARY

### NAVIGATION: NAV AUTO COMMANDS (L3)
### SUBSET M (10 COMMANDS)

10. BY
11. CAPTURE
23. DEGRESS
36. FLY TO
50. INITIALIZE
57. MERGE
76. SECTOR
87. TIME
88. TRACK
98. WAYPOINT

98.72% CORRECT
RANK = 8

### NAVIGATION: ARRIVAL TIME COMMANDS (L3)
### SUBSET N (10)

3. AIRSPEED
6. AUTO
20. COUPLE
25. DISPLAY
29. EMERGENCY
42. GO AROUND
68. OVER
78. SET
91. UNDER
93. VECTOR

98.43% CORRECT
RANK = 12

### NAVIGATION: GENERAL COMMANDS (L4)
### SUBSET O (10)

9. BLANK
13. CHANGE
17. CLEAR
31. ENTER
32. ERROR
33. EXECUTE
41. GO
61. NEGATIVE
71. POSITIVE
95. VERIFY

98.47% CORRECT
RANK = 11

TABLE 9

## RANK ORDER OF SUBSETS: 100-COMMAND VOCABULARY

SUBMAX$_{(25)}$ – NO REJECTION

| SUBSET | MEAN PERCENTAGE CORRECT BY SUBSETTING | RANK BY SUBSETTING | RANK OVERALL | MEAN PERCENTAGE CORRECT OVERALL |
|---|---|---|---|---|
| J (3) | 99.70 | 1 | 2 | 95.56 |
| I (3) | 99.53 | 2 | 1 | 95.60 |
| B (4) | 99.15 | 3.5 | 5 | 94.48 |
| G (6) | 99.15 | 3.5 | 12 | 92.45 |
| E (6) | 99.03 | 5 | 8 | 93.67 |
| H (5) | 98.86 | 6 | 14 | 91.12 |
| C (4) | 98.725 | 7 | 3 | 95.53 |
| M (10) | 98.720 | 8 | 10 | 92.57 |
| A (6) | 98.70 | 9 | 4 | 94.75 |
| F (9) | 98.48 | 10 | 6 | 94.33 |
| O (10) | 98.47 | 11 | 9 | 93.56 |
| N (10) | 98.43 | 12 | 7 | 94.03 |
| L (4) | 98.40 | 13 | 11 | 92.55 |
| D (10) | 98.21 | 14 | 13 | 91.79 |
| K (10) | 97.98 | 15 | 15 | 91.02 |

Recognition accuracy and rank order results for the 10 digits are shown in Table 10. Recognition accuracy was highest for "six" and lowest for "five" by both subset and overall processing. Among individual commands in the entire 100-command vocabulary, "six" ranked 11.5 in recognition accuracy by subsetting and 2nd by overall processing. For comparison, "smaller" (Subset J) ranked 1st by subsetting (37th by overall) with 99.9% correct, and "waypoint" (Subset M) ranked 1st by overall processing (11.5) by subsetting) with 98.3% correct. In contrast, the digit "five" ranked 100th by both subsetting and overall processing. Thus differences in recognition accuracy within the digit subset cover nearly the entire range of results obtained over all 100 individual commands.

Within the digit subset, the most of the recognition misclassifications in subset processing occurred between "five" and "nine". Of the total misclassifications of "five" nearly 80% had been recognized as "nine", while more than 80% of all misclassifications of "nine" had been recognized as "five". For the digit subset, recognition accuracy could be improved by having speakers pronounce "nine" as "niner" (they did not do so in this study). Similar pronunciation changes could be made in other subsets where misclassification between two commands accounts for a high percentage of the total error. In some cases, a new command should be substituted for one of the problem commands and the evaluation process repeated.

The mean percentage difference in recognition accuracy between subset and overall processing for each digit is shown in Table 11. The relative cost of attempting to recognize a digit while protected within its own subset vs. leaving it unprotected in the overall processing is indicated by the value shown for each digit.

New vocabularies proposed for flight systems use may be evaluated by the process described for evaluation of the 100-command vocabulary. Thus any serious incompatibilities between commands within a subset may be identified and corrected, and recognition accuracy for the entire vocabulary can be determined prior to actual use in a flight system.

Results of the 100-command vocabulary evaluation indicate that recognition accuracy of the Ames speech recognition system is sufficiently high to permit operational testing.

## OPERATIONAL TESTING IN NOISE AND VIBRATION

Since both noise and vibration are present in flight environments and since both threaten to reduce the high levels of recognition accuracy obtainable under laboratory conditions, a system recognition accuracy evaluation is scheduled for several different conditions of noise or vibration.

TABLE 10

## RANK ORDER OF DIGITS: 100-COMMAND VOCABULARY

SUBMAX$_{(25)}$ – NO REJECTION

| DIGIT | MEAN PERCENTAGE CORRECT BY SUBSETTING | RANK BY SUBSETTING | RANK OVERALL | MEAN PERCENTAGE CORRECT OVERALL |
|---|---|---|---|---|
| SIX | 99.5 | 1 | 1 | 98.0 |
| THREE | 99.3 | 2 | 5 | 94.2 |
| ZERO | 99.8 | 3 | 2 | 96.7 |
| FOUR | 98.7 | 4.5 | 4 | 95.2 |
| SEVEN | 98.7 | 4.5 | 7 | 88.2 |
| ONE | 98.2 | 6.5 | 6 | 94.0 |
| TWO | 98.2 | 6.5 | 3 | 96.3 |
| EIGHT | 97.8 | 8 | 9 | 85.5 |
| NINE | 96.6 | 9 | 8 | 87.8 |
| FIVE | 96.3 | 10 | 10 | 81.9 |

TABLE 11

## PERCENTAGE DIFFERENCES FOR DIGITS: 100-COMMAND VOCABULARY

### SUBSET vs. OVERALL PROCESSING

| DIGIT | MEAN PERCENTAGE DIFFERENCE |
|---|---|
| SIX | 1.5 |
| TWO | 1.9 |
| ZERO | 2.1 |
| FOUR | 3.5 |
| ONE | 4.2 |
| THREE | 5.1 |
| NINE | 8.8 |
| SEVEN | 10.5 |
| EIGHT | 12.2 |
| FIVE | 14.4 |

To provide a more realistic test of the Ames ASR system than would be obtained if speakers simply voiced commands in a noise or vibrating environment, voice command data will be collected while the test subjects perform a continuous tracking task. Comparable keyboard entry data will also be collected.

## Background

During flight a pilot must perform a variety of tasks, most of which may be assigned to one of two broad categories: (1) tracking, e.g., following a glideslope or making a turn, and (2) system interaction, e.g., setting an autopilot heading or changing a radio frequency. The use of digital computers on board aircraft (primarily for automatic subsystems control) is already a reality, and in future aircraft the role of the computer will be large [1,5]. Thus tasks that involve interacting with onboard avionics systems are increasingly becoming tasks of man-computer interaction. Effective utilization of the computer's information processing capabilities requires careful study and critical testing to provide design guidelines for the man-computer interface. In particular, a crew member must be able to provide inputs in a manner that is

(1) accurate,
(2) tolerant of errors and updates,
(3) rapid enough to meet the demands of the task at hand,
(4) natural and convenient, so that use of the input system does not add significantly to the user's workload, and
(5) interruptable.

Since speech has the potential for equaling or exceeding conventional keyboard input capabilities in meeting these requirements [6], the operational testing is designed to assess the relative effectiveness of voice and keyboard input systems in stationary, noisy, and vibrating environments.

## Plan of Study

Pilots will act as test subjects and will perform a single-axis compensatory tracking task while being exposed to various levels of noise. While tracking, the pilots will concurrently make discrete numerical data entries, upon command, using one of two input media: voice or keyboard. Following completion of the required number of noise evaluation test sessions, the pilots will perform the same tasks for an equivalent number of test sessions under various levels of vibration. The pilots will be selected from the general aviation, military, and airline pilot populations.

All noise and vibration test data will be collected in the Ames Vertical Acceleration and Roll Device (VARD). The noise and vibration will simulate conditions occurring on board aircraft. Four noise conditions will be tested:

(1) No Noise,
(2) Helicopter noise at 90 dB,
(3) Helicopter noise at 100 dB,
(4) Random ("white") noise at 100 dB.

For each pilot, no vibration testing will be imposed until all noise testing has been completed. Four vibration conditions will then be simulated and tested:

(1) No vibration,
(2) Smooth jet transport cruise,
(3) Rough jet transport cruise,
(4) Helicopter cruise.

For all test conditions, the primary performance measures will be tracking error and the accuracy and speed of voice and keyboard data entries. Assessment of the relative effect on tracking performance of making voice vs. keyboard entries under the various noise and vibration conditions is of particular interest.

Depending upon initial noise and vibration test results, additional evaluation may be desirable for other noise or vibration conditions, or combinations of noise and vibration. When favorable ASR system results have been achieved under the noise and vibration conditions expected in flight, the system will be ready for flight testing.

## FLIGHT APPLICATIONS

The hardware requirements for a flyable ASR system are quite different from the requirements for a laboratory system. Selection of a hardware system that will be fully operative under a variety of flight conditions is essential for later success in flight.

### The Ames ASR Flight System

The primary component of this Ames ASR system is a Rolm 1603A ruggedized military computer. The entire system is contained in a 7.6" x 10.1" x 20.4" chassis and weighs about 60 lbs. The system meets military physical operating standards set by MIL-E-5400 Class II and MIL-E-16400 Class I specifications with respect to operating temperature, vibration, shock, humidity, altitude, and radio frequency interference. It contains 32,764 16-bit words of memory, central processor with extended arithmetic unit, several general purpose input/output ports, and a microprocessor-based speech input board. This board accepts microphone inputs and performs shaping, spectral analysis, word boundary detection, nonlinear time normalization, and coding completely independent from the 1603A processor. The 1603A accepts 128-bit patterns from the speech input board, performs training and recognition functions, and directs overall control of a flight experiment. For a single user, structured vocabularies of up to 200 words may be accommodated.

## Potential Flight Applications

Although the potential flight applications of ASR systems are numerous, those applications that seem particularly desirable have general characteristics that would allow the ASR system to (1) reduce the difficulty (and risk) of performing high-wordload manual control procedures during critical phases of flight, (2) minimize eye-hand coordination problems that often accompany a heavy manual control burden, and (3) reduce visual time-sharing requirements so that attention can remain focused outside the cockpit or upon a primary flight display for longer periods of time. For example, the selection and tuning of radio frequencies, a procedure that often interferes with other essential manual tasks, could be controlled by voice. Computer generated displays, including head-up displays, could be controlled by voice rather than by conventional keyboard, thus avoiding eye-hand coordination problems and allowing visual attention to remain focused on the display. In addition, whenever command sequences or numerical data entries must be (1) performed manually, (2) integrated with other manual tasks, and (3) executed rapidly, e.g., a Navy P-3 sensor operator's use of thumbwheel switches for numerical data entry, voice commands have the potential for providing faster and more accurate performance, as well as being less disruptive of (and less disrupted by) other manual task requirements [7].

Once selection of a specific ASR flight application has been made and all hardware and software requirements have been met, full-mission flight simulation testing is desirable prior to actual use in flight.

Several candidate ASR flight applications are being considered for in-flight evaluation of the Ames ASR Flight System. Three of these potential applications have been selected for discussion.

## Navy P-C3 Orion Pilot Application

During certain phases of anti-submarine patrol flights, Navy P-3C aircraft must be flown at low altitude. While low altitude flight is maintained, the pilot must select and execute command functions using a 35-key keyset that is located to his right and slightly behind him. This keyset location requires that the pilot turn his body towards the keyset and direct his visual attention away from the outside visual scene and his primary flight instruments while using the keyset. Use of the Ames ASR Flight System, with a 34-command syntax-structured vocabulary consisting of 11 subsets ranging in size from 2 to 14 commands would provide a command capability duplicating that of the keyset and would increase the pilot's capability for rapid detection and correction of significant deviations from the desired flight conditions during this time-critical phase of flight.

Lt. Anthony Quartano, a Navy P-3C pilot, developed the procedure used for selection of the voice command vocabulary and designed the command syntax in collaboration with the authors of this paper.

The speech recognition system would be implemented in parallel with the pilot keyset in a P-3C aircraft for in-flight evaluation of voice command system performance; thus comparable speed and accuracy data could be collected by voice and by keyset, and access to normal keyset functions would be available at all times.

## Navy P-3C Sensor Station Operator Application

Sensor Station 1 and 2 operators on board Navy P-3C aircraft use several different keysets and thumbwheel switch input sets to enter data into a digital computer. During some missions, the desired rate of information entry greatly exceeds the rate obtainable with current input devices. Several Navy P-3C sensor operators based at Naval Air Station, Moffett Field, CA, have collaborated with the authors of this paper to evaluate potential ASR sensor station applications. Five specific applications were selected that would reduce an operator's manual workload and permit more rapid execution of essential tasks. For example, use of voice commands rather than thumbwheel switches for numerical data entry would increase the maximum entry rate while reducing interference with other concurrent manual tasks. Implementation of voice command functions would be made in parallel with existing input hardware so that comparable in-flight data could be collected by voice and by the conventional manual input method, and to provide access to normal hardware functions at all times.

Sufficient space is available in the Ames ASR Flight System chassis to accommodate up to four speech input boards, allowing the system to simultaneously process information from four different users. These users may or may not share vocabularies and/or input functions. This multi-user capability is made possible by the microprocessor-based speech input boards which relieve the 1603A computer of a large percentage of the total processing load. Thus the Ames ASR Flight System could easily process information in parallel from both Sensor Stations 1 and 2 during flight.

## Helicopter Pilot Application

Helicopters, with their high noise and vibration characteristics, provide the most challenging and perhaps one of the most deserving flight environments for ASR applications. Because the manual control burden in helicopters is typically quite heavy, the implementation of an on-board ASR system could provide substantial benefits for increasing the efficiency and safety of helicopter flight. For example, during helicopter missions where success may require operating the aircraft near its operational boundaries, e.g., in search, rescue, or evacuation missions, the pilot could benefit from knowing how close he is to exceeding critical system performance limits such as altitude, airspeed, or gross weight. Given that an appropriate computational capability exists on the aircraft, information such as hover ceilings, gross weight margins, engine performance parameters, and range-of-flight estimates could be presented to the pilot in response to his voice commands. The fact that in these situations the pilot is in a high-workload manual control environment with inherent time-stress recommends the application of speech technology. An ASR system would not only allow the pilot to conveniently query the computational system for specific critical information, but speech synthesis could also be used to minimize the need for diversion of attention to a conventional visual display.

At the present time, Ames Research Center has assumed the lead role for NASA helicopter research, and applications such as this are being actively pursued. Fortunately, a variety of helicopters will be available for flight research in the near future, and in-house avionics groups are working on developmental projects such as the example discussed above. Accordingly, a very exciting and rewarding future is anticipated for this kind of man-machine systems research.

## REFERENCES

1. Wempe, T.E. Flight Management--Pilot Procedures and System Interfaces for the 1980's - 1990's. AIAA Paper No. 74-1297, Amer. Inst. Aero. and Astro. Life Sciences and Systems Conference, Arlington, Texas, November 1974.

2. Nilsson, N.J. Learning Machines. New York: McGraw-Hill, 1965.

3. Hillborn, E.H. Keyboard and Message Evaluation for Cockpit Input Data Link. Report No. DOT-TSC-FAA-71-21, Dept. of Transportation, Washington, D.C., 1971.

4. Coler, C.R. and Plummer, R.P. Development of a Computer Speech Recognition System for Flight Systems Applications. Aersopace Medical Association, Preprints of the 45th Annual Scientific Meeting, Washington, D.C., May 1974, Pp. 116-117.

5. Belson, J. The 21st-Century Flight Deck. Flight International, April 23, 1977, 111, Pp. 1118-1120.

6. Turn, R. Speech as a Man-Machine Communication Channel (Report No. P-5120). Santa Monica: Rand Corp., January 1974.

7. Plummer, R.P. and Coler, C.R. Speech as a Pilot Input Medium. Proceedings of the Thirteenth Annual Conference on Manual Control, Cambridge, Mass., June 1977, Pp. 460-462.

## DISCUSSION

## Dr. Edward Huff

Q: Jerry Wolf, BBN: In one of your slides, Clay, you presented some recognition results for the entire 100-word vocabulary and then for comparison the 10 digits. The 10 digits came out with a smaller recognition score than the entire 100-word vocabulary. That seems a little inconsistent to me. I don't see how the smaller vocabulary could have come out with a significantly worse score than the 100. Do you have any comments on why that happened?

A: Clay Coler: As I indicated when we went through the results, when individual words were recognized over the entire vocabulary, some words had quite high recognition accuracy. For example, the digit six was recognized almost as well in the overall processing as it was by subsetting. Other digits were quite poor, and in fact the digit five was the poorest of all 100 commands. I think that the answer is simply that the digits are relatively tough, no matter whether they are protected in a subset or left unprotected in the overall processing.

Q: Jerry Wolf: They are significantly worse than the recognizability of that whole 100. They were relatively distinguishable.

A: Clay Coler: They are certainly significantly worse than other 10-command vocabularies and we can see that when we artificially put subsets together in the overall processing.

Bob Plummer: Maybe it wasn't clear that the ten words were still recognized against the whole hundred, but he was showing you the result of those ten particular words, the digits. So, that's how thay could come out with a lower score.

Q: Mike Grady, Logicon: I have to say that the work the NASA has done on trying to assess recognition accuracy is probably the most extensive I've seen, and I really want to express to all of you appreciation for that. I do have some questions on the experimental design or whatever you would like to call it. I guess I'll address this to you, Bob. In your comparison with 20 speakers in the six different training recognition strategies that you used, did you use the same 20 people when you went through each of the strategies, and also did the speakers, when they were performing the experiment, get feedback on the results that they were getting? I might point out that the reason I'm asking this question is because we found that there is really a phenomenon of learning how to talk to the boss, and it's a thing that we've never looked at and I'm curious about your results.

A: Bob Plummer: The idea there was to examine recognition algorithms separately from the frontend or the initial digitization. So, all the speaker did was say the right word at the right time. We would then digitize that and record the digital version of the word on tape. We did not do recognition at all on line. So the speaker received absolutely no feedback as to whether he was right or wrong. I mean for better or for worse, he didn't get it. What that gives us may be slightly unrealistic in the sense that in an application he probably would get feedback. On the other hand, it means that we have absolute repeatability because we can just cram those digitized words into any different recognition algorithm that we might want to use. So we really were applying all the different algorithms to exactly the same set of digitized data.

Q: John Markel Signal Technology: You talk about having 10-word reference or twenty-five words. Would you describe your training scenario? Whether that's taken at one setting or spaced over time. How is that done?

A: Bob Plummer: These particular tests were all done in the following way. The subject would see on a display the word or digit he would be asked to pronounce at a given time. All these data were taken sequentially, so in his first session he would provide us with twenty-five training samples of each word, and we would do that by cycling through the vocabulary, so he didn't say "waypoint" 25 times and then "altitude" 25 times. He would cycle through 25 times. Then after perhaps a short rest, he would go right on into the recognition. This all took place in one day, and more or less continuously through time. So we weren't really facing problems of day to day variation. Although I guess actually that is not completely true. Some of the data collection did take place over several days. Let me ask Clay--how many days were involved?

Clay Coler: For the 10-word vocabulary, all data were collected in a single day. For the 100-command vocabulary, it was five days. The first day was the training data day.

Bob Plummer: O.K., so what I said about the one day was right for the 10 words. For the 100 words, it was over five days. Training on the first day and recognition on the next four.

Q: Marv Herscher, Threshold Technology: I noticed on the slides and tables of the digits subsets, I think three times they appeared. One originally and two as subset-in your test, and there seemed to be some inconsistency in terms of the accuracies that were posted among the three. Would you comment on how this comes about?

A: Bob Plummer: I think that might be related to the original question It might be better to have a look later, privately or in the published paper. The question has come up in the following way: We were recognizing some times perhaps against only ten words, and that's what we call subsetting. So there the 10 digits were recognized only against other digits. At other times, we were recognizing overall. The digits were recognized against all 100 words but we artificially extracted out of that the digit subset to see how well those 10 words did as a group when compared to all 100 words. So those two results might look kind of different, but in fact I think we're consistent with what happened overall.

Q: Marv Herscher: I don't understand because I thought you said that the second column you had on a couple of the tables were digit subsets against each other. A true subset is a subset of words recognized only against each other.

A: Bob Plummer: Right. It was done three ways. The first was only the digit vocabulary. Then there was the big vocabulary with the digits as a true subset. And a big vocabulary with digits versus everything. Would you like me to put the slide back up?

Marv Herscher: Well, I can talk to you about it later.

[redacted]

# VOICE DATA ENTRY IN AIR TRAFFIC CONTROL

DONALD W. CONNOLLY

NATIONAL AVIATION FACILITIES EXPERIMENTAL CENTER
ATLANTIC CITY, NEW JERSEY

## BACKGROUND

The introduction of large-scale automation into civil air traffic control is relatively recent. Until 1970 there were only a very few isolated and largely developmental installations. Since 1970 all 20 of the enroute control centers and 64 of the major terminal control facilities have had large, computer-based systems installed and commissioned for operation. These systems function in many ways analogously to military command, control, information and communication systems and they are all directly or indirectly interconnected.

The whole air traffic control complex is basically a cooperative surveillance and control operation. Its functioning depends on many elements, not the least of which is complete, correct and up-to-date flight plan and flight progress information. The execution of flight plans precisely as filed in advance, however, is more exceptional than routine. Even "standard" airline plans for scheduled flights are subject to change before departure as well as while enroute due to many factors, most notably weather and wind conditions. The air traffic control specialist, of course, is the principal point of direct contact between the traffic management system and the traffic itself and is a major conduit of information between them. While the function of the system is to maintain and provide vital information to the controller, he or she in turn has the task of supplying a substantial quantity of information to the system.

Automation has altered a number of task elements of the job of the air traffic controller. In most instances these changes have been in the direction of improved quality, efficiency and simplicity though they have by no means diminished the complexity or responsibility of the job of traffic controller. While some types of workload have been reduced or nearly eliminated (visual/manual tracking, maintenance of identification, acquisition/maintenance of altitude information) new tasks have been added. Perhaps the most significant and onerous of the latter is that of manual entry of flight data and system commands and queries. As in many computer-based operations the "language" used for entry or query is an abbreviated, partially mnemonic, coded language. Even so, the key-entry workload and its potential for distraction and interference with the main stream of the controllers' task, remains high. In peak traffic hour, for example, an enroute traffic controller will

[redacted]

commonly find it necessary to enter messages into the system computer which aggregate to 700 or more single keystrikes.

Human factors and air traffic control specialists at the National Aviation Facilities Experimental Center have recognized and been concerned with information transfer problems at the controller-computer interface throughout the system development process. A major interest, of course, has been the area of data entry and system control. Emerging technologies of information presentation and data input are continually reviewed and promising techniques experimentally investigated. A number of variations of the "touch" or "menu-select" principle of "chunk" data entry (as versus character-by-character message composition) have been tried in laboratory and simulation experiments, for example. We have been aware of, and following, the development of spoken word recognition technology since at least 1971. It was not until the middle of 1975, however, that we were able to secure the approval and resources necessary to undertake in-service exploration of the applications of word recognition (and, by some logical and temporal extension, speech understanding) in the field of air traffic management. Thus far the magnitude of effort underway in the Federal Aviation Administration has been rather small (one scientist with part-time aid of one technician and one programmer) and has been directed toward application, adaptation and modification of speech recognition technology rather than development of the technology itself.

## PROGRESS

### Introduction

In May 1975, a basic Threshold Technology, Inc., model VIP-100 was acquired for use in a series of word-recognition applicability studies. This equipment included an ASR-33 Teletype, a NOVA 2 minicomputer with 16K of core memory, the Threshold digitizer and a three-transport cassette tape unit. A Tektronix model 4012 CRT/keyboard computer terminal was added as the basic output device. At various times since 1975 additional hardware has been secured, including a 10 megabyte disk store, a Digital Equipment Corp. DECwriter, 16K words of core memory and an in-house designed and fabricated voice digitizer whose uses and potential uses will be described below under post FY-77 efforts. The equipment of the Voice Entry Laboratory is shown schematically in Figure 1.

Several of the keyboard data entry "languages" of the National Airspace System were tabulated and analyzed. There are two such languages in regular and extensive use in the semi-automated enroute traffic control centers of the agency which produce daily hundreds of thousands of messages requiring millions of keystrikes. There are a number of

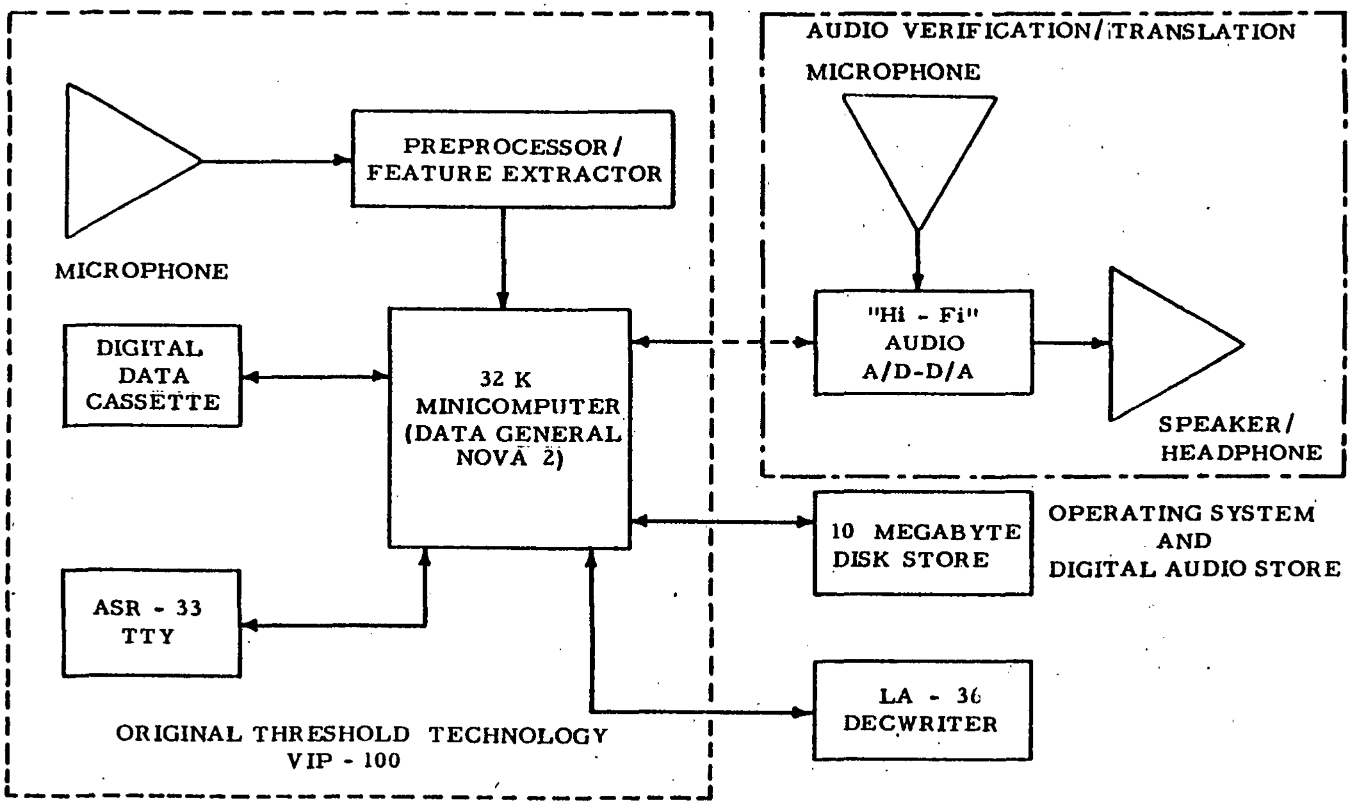

MICROPHONE
PREPROCESSOR / FEATURE EXTRACTOR
DIGITAL DATA CASSETTE
32 K MINICOMPUTER (DATA GENERAL NOVA 2)
ASR - 33 TTY
ORIGINAL THRESHOLD TECHNOLOGY VIP - 100
AUDIO VERIFICATION/TRANSLATION
MICROPHONE
"Hi - Fi" AUDIO A/D-D/A
SPEAKER/ HEADPHONE
10 MEGABYTE DISK STORE
OPERATING SYSTEM AND DIGITAL AUDIO STORE
LA - 36 DECWRITER

FIGURE 1
FAA/NAFEC VOICE ENTRY LABORATORY

other entry languages in the system (e.g. control tower cab, terminal radar control facility, flight service station, etc.) which are either not as burdensome or distracting, or not as complex and voluminous in use, or both, but which are also likely candidates for application of word recognition technology. The key language which was chosen as the test vehicle was that used by the non-radar or flight data controllers in enroute control centers. The structure and vocabulary of this langusage may be found in Tables 1 and 2. This particular language was selected for a number of reasons. In the first place, it is one of the more complex languages in use. The total repertoire of possible messages is larger than that of any of the other key languages used by personnel engaged in the active control of traffic. Finally, the key-entry workload at this operational position is the largest in total volume in the system. Thus, a very difficult application was undertaken for investigation right at the outset. The theory behind this choice was that (a) it appeared highly likely, given the state of the word recognition art, that this application would be practical and cost/beneficial and that, a fortiori, less complex, less difficult applications would yield to the same approach with zero or minimum additional research and development effort or that (b) many or most of the relevant questions for the lesser applications would be answered in the course of attacking the greater, even if the present state of technology did not prove practical for this particular application.

## Initial Experiments

The language chosen for test was found to include a total of 24* basic types of messages. Of these, 15 types of messages encompass 96 percent of all messages actually entered in operation. In addition, these 15 message types include all of those occurring with a frequency of one in a hundred or greater. The first element of every message is the message type. It was also found that, in most cases, the type of message must be followed by the identity of the flight data file (flight plan) to which the entry applies. Furthermore, of the four means of identifying a flight, the one most commonly employed was the three-decimal-digit computer identity number assigned to every flight. Thus, the second element of most spoken messages could be assembled from a word list consisting only of digits plus two or three control words (such as "erase" for restarting the whole entry and "backspace" for changing the last digit.)

*An additional seven types of message (covering "conflict alert" entries) have since been added. This, based on experience to date, should not cause any special difficulty.

## TABLE 1

VOICE DATA ENTRY: D-CONTROLLER LANGUAGE STRUCTURE

| KIND OF MESSAGE | | | | |
|---|---|---|---|---|
| AMEND | 3 DIG IDENT | DATA FIELD NAME | DATA ENTRY FOR FIELD | GO |
| CORRECTION | | " | " | " |

| DATA FIELD NAMES | VOICE ENTRIES REQUIRED |
|---|---|
| Type | See Below |
| Beacon Code | 4 Octal Digits |
| Speed | 3 Decimal Digits |
| Fix | Place Name |
| Time | 4 Decimal Digits |
| Altitude | 3 Decimal Digits |
| Qualifier | See List of Qualifiers and Note 2 |
| Ident | 7 Alphanumerics (decimal) |

Note 1: After a "field name" and appropriate entries for that field have been entered the system will accept another field name (plus proper entries) and yet another etc. without limit OR it will accept an ERASE command, a BACKSPACE command or a GO (ENTER) command.

FOR "TYPE" ENTRIES, ALWAYS SAY:

| | | | |
|---|---|---|---|
| | MFG NAME, | 2 or 3 A/N, | Name a Qualifier |
| or | MILITARY, | 4 A/N, | " " " |
| or | GENERAL, | 4 A/N, | " " " |

| IF YOU SAY: | YOU'LL SEE: | THEN SAY: |
|---|---|---|
| Boeing | B | 3 A/N e.g. 707 |
| British | BA | 2 A/N e.g. 11 |
| Vickers | VC | 2 A/N e.g. 10 |
| Lockheed | L | 3 A/N e.g. 011 |
| Nord | N | 3 A/N e.g. 026 |
| Dehavilland | DH | 2 A/N e.g. C6 |
| Douglas | DC | 2 A/N e.g. 10 |
| Military | --- | 4 A/N e.g. C131 |
| General | --- | 4 A/N e.g. PA13 |

TABLE 1 (Continued)

TO ENTER A "TYPE, YOU MUST ALWAYS ADD ONE OF THE EQUIPMENT QUALIFIERS:

| IF YOU SAY: | YOU'LL SEE |
|---|---|
| Discrete | /U |
| DiscreteDME | /A |
| DME | /D |
| Nondiscrete | /T |
| NondiscreteDME | /B |
| Transponder | /X |
| TransponderDME | /L |
| TACAN | /M |
| TACAN64 | /N |
| TACANDiscrete | /P |

Note 2: If you wish to enter an amendment to the QUALIFIER part of the "type" field alone, you need only name the data field "QUALIFIER" then name one of the qualifiers above.

FINALLY, YOU MAY SAY "GO" (to ENTER), BACKSPACE (if you wish to change or correct an error of entry or of recognition) or "ERASE", OR, YOU MAY NAME ANOTHER DATA FIELD AND CONTINUE AS BEFORE.

| KIND OF MESSAGE | SEQUENCE | | |
|---|---|---|---|
| REPORTALTITUDE | 3 DIG IDENT, | 3 DECIMAL DIG (ALT), | GO |
| DISCRETECODE | " | 4 OCTAL DIG (CODE), | GO |

These are shorthand messages requiring only the KIND name, the track I.D. and the data to be entered. It is realized that discrete codes are often assigned automatically upon entry request in NAS. The voice entry here is for test purposes.

TABLE 1 (Continued)

| KIND OF MESSAGE | SEQUENCE | |
|---|---|---|
| DROPTRACK | 3 DIG IDENT, | GO |
| PRINTSTRIP | " | " |
| ACCEPTHANDOFF | " | " |
| READOUT | " | " |
| CANCEL | " | " |

These messages are all identical excpet for the first word, the kind of message.

| KIND OF MESSAGE | SEQUENCE | | | |
|---|---|---|---|---|
| DEPARTURE | 3 DIG IDENT, | 4 DEC DIG (TIME) | NAME (FIX) | GO |
| HOLD | " | 4 DIG (TIME) | NAME (FIX) | GO |
| RELEASE | " | 4 DIG (TIME) | | GC |
| TRANSMIT | " | NAME (FIX) | | GO |

These messages require entry of a four digit time, or a one word place name (FIX) or both in addition to the message kind and the identity of the flight.

| KIND OF MESSAGE | SEQUENCE | | |
|---|---|---|---|
| WEATHER | NAME(FIX) | | GC |
| HANDOFF | 2 DIG (SECTOR) | 3 DIG IDENT | GC |

These and CORRECTION (above) are the only kinds of messages that are not immediately followed by identity.

TABLE 2

VOICE DATA ENTRY: D-CONTROLLER VOCABULARY

| WORD NO. | UTTERANCE | PRINT WORD DISPLAYED |
|---|---|---|
| DIGITS | | |
| 0 | ZERO | 0 |
| 1 | ONE | 1 |
| 2 | TWO | 2 |
| 3 | THREE | 3 |
| 4 | FOUR | 4 |
| 5 | FIVE | 5 |
| 6 | SIX | 6 |
| 7 | SEVEN | 7 |
| 8 | EIGHT | 8 |
| 9 | NINER | 9 |
| CONTROL WORDS (SEE ALSO #102 ERASE) | | |
| 10 | BACKSPACE | |
| 11 | GO | (ENTER) |
| MESSAGE TYPES | | |
| 12 | AMEND | AM |
| 13 | CANCEL | CN |
| 14 | CORRECTION | CR |
| 15 | DEPARTURE | DM |
| 16 | DISCRETECODE | DQ |
| 17 | READOUT | FR |
| 18 | ACCEPTHANDOFF | HO |
| 19 | HANDOFF | HO |
| 20 | DROPTRACK | RS |
| 21 | PRINTSTRIP | SR |
| 22 | HOLD | HM |
| 23 | RELEASE | HM |
| 24 | REPORTALTITUDE | RA |
| 25 | WEATHER | WR |
| 26 | TRANSMIT | XM |
| FLIGHT DATA FIELD NAMES | | |
| 27 | TYPE | 03 |
| 28 | QUALIFIER | 03 |
| 29 | BEACONCODE | 04 |
| 30 | SPEED | 05 |
| 31 | FIX | 06 |
| 32 | TIME | 07 |
| 33 | ALTITUDE | 08 |
| 34 | IDENT | 02 |

TABLE 2 (Continued)

### FIXES

| | | |
|---|---|---|
| 35 | WILLIAMSPORT | IPT |
| 36 | SELINGSGROVE | SEG |
| 37 | MILTON | MIP |
| 38 | HAZELTON | HZL |
| 39 | WILKESBARRE | AVP |
| 40 | EASTTEXAS | ETX |
| 41 | LAKEHENRY | LHY |
| 42 | TOBYHANNA | TSD |
| 43 | ALLENTOWN | ABE |
| 44 | STILLWATER | STW |
| 45 | BENTON | 7QB |
| 46 | SWEETVALLEY | 7EV |
| 47 | LOPEZ | 7LE |
| 48 | SNYDERS | 7YX |
| 49 | SLATINGTON | 7ZO |
| 50 | WHITEHAVEN | 9WT |
| 51 | RESORT | 9ZT |
| 52 | PENNWELL | 7PW |
| 53 | HUGUENOT | HUO |
| 54 | SOLBERG | SBJ |
| 55 | FREELAND | 7FE |

### AIRCRAFT TYPE NAMES

| | | |
|---|---|---|
| 56 | BOEING | B |
| 57 | DOUGLAS | DC |
| 58 | LOCKHEED | L |
| 59 | CONVAIR | C |
| 60 | VICKERS | VC |
| 61 | NORD | N |
| 62 | BRITISH | BA |
| 63 | GENERAL | - |
| 64 | MILITARY | - |
| 65 | DEHAVILLAND | DH |

### PHONETIC ALPHA

| | | |
|---|---|---|
| 66 | ALPHA | A |
| 67 | BRAVO | B |
| 68 | CHARLIE | C |
| 69 | DELTA | D |
| 70 | ECHO | E |
| 71 | FOXTROT | F |

## TABLE 2 (Continued)

PHONETIC ALPHA (Continued)

| | | |
|---|---|---|
| 72 | GOLF | G |
| 73 | HOTEL | H |
| 74 | INDIA | I |
| 75 | JULIET | J |
| 76 | KILO | K |
| 77 | LIMA | L |
| 78 | MIKE | M |
| 79 | NOVEMBER | N |
| 80 | OSCAR | O |
| 81 | PAPA | P |
| 82 | QUEBEC | Q |
| 83 | ROMEO | R |
| 84 | SIERRA | S |
| 85 | TANGO | T |
| 86 | UNIFORM | U |
| 87 | VICTOR | V |
| 88 | WHISKEY | W |
| 89 | XRAY | X |
| 90 | YANKEE | Y |
| 91 | ZULU | Z |

"QUALIFIERS"*

| | | |
|---|---|---|
| 92 | DISCRETE | /U |
| 93 | DISCRETE DME | /A |
| 94 | DME | /D |
| 95 | NONDISCRETE | /T |
| 96 | NONDISCRETE DME | /B |
| 97 | TRANSPONDER | /X |
| 98 | TRANSPONDER DME | /L |
| 99 | TACAN | /M |
| 100 | TACAN 64 | /N |
| 101 | TACAN DISCRETE | /P |

*These expressions are to be said as all one word such as "discrete dee em ee", even though printed here and on the training display as separate words.

CONTROL WORD
(SEE ALSO #10 BACKSPACE AND #11 GO)

| | | |
|---|---|---|
| 102 | ERASE | Erases Entry |

The second element of some types of messages (e.g., weather information retrieval) and third or fourth element of other messages (e.g., early handoff to a terminal; hold message) is a location identifier or geographic "fix." The keyboard codes for these place names are not always mnemonic (e.g., Benton is coded 7QB) but the place names themselves are easily spoken. No attempt was made to survey all possible fix-names; however, the list included for one sector in the New York ARTCC, all VOR's, all intersections, and all terminals; in short, all the fixes normally required at the position as elements of key-entry messages.

Two types of messages (flight plan amendment and correction thereof) require identification or naming of a flight plan data field (e.g., assigned altitude; speed). Eight of these data fields account for the vast majority of modifications entered and the field content or substantive data most commonly consist of digits.

Certain types of entries or, more precisely, parts of messages currently made with keyboards basically exist only in coded, nonverbal or partially nonverbal form. Consider the aircraft identity N1009Y (tail number). The most convenient way to make such an entry might still be via keyboard. However, an "all purpose" subvocabulary consisting of all of the digits plus the phonetic alphabet (which is part of the linguistic stock-in-trade of the traffic controller) were made a part of the total vocabulary of the voice data entry language for the purpose of making the comparatively fewer and rarer entries not already encompassed by the word lists described above.

These subvocabularies, plus a short list of commercial aircraft types and the list of relevant avionics equipments (or type "Qualifiers"), make up the whole vocabulary as currently constituted. The vocabulary and syntax of the language, as previously noted, are included here as an appendix.

The first experiments which were conducted were intended to establish the basic recognition performance of the VIP-100 word recognition package with three of the subvocabularies discussed above, namely the 15 message types word list, the 21 fix names list, and the 10 digits (plus "erase" and "backspace") list. Each of the lists, separately was expanded into a pseudo random assembly in which each member of the list appeared 10 times. Thus the "reading list" for message types was 150 "words" long, that for "digits" 120 words, and for fixes, 210 words. Each speaker then "trained" the word recognizer by speaking each expression (some, as may be seen in the appendix, were composites or phrases spoken without internal pauses) 10 times. This resulted in composite digital images of the way the speaker speaks each of that particular list of words. These reference images were then written on cassette tape for later reuse. Following the initial "training" session, each speaker reads the random list described above on 10 separate occasions in the case of message types and fixes, 5 sessions for the digits list. Data were automatically collected during each test session on the number of times each word was correctly recognized, the number of times incorrectly recognized, the average closeness of match between the spoken entry and the best and second-best choice among the reference images (i.e., the training images), and the duration of the spoken expression. Each subject, over a period of several days to several weeks, spoke (for recognition testing) each word in each of the subvocabularies 100 times for the types and fixes and 50 times for the digits. The principal purpose of testing digits at all was to ascertain whether our sample of speakers produced the order of recognition accuaracy for digits which is commonly found using this word recognition equipment.

## Initial Results

A total of 12 speakers served as test subjects for Phase I. Nine were journeyman air traffic control specialists with extensive experience in the National Airspace System Enroute Test Facility. Three were non-controllers, two female and one male. No differences were found that could be attributed to either profession or gender. One group of 11 of these speakers served as subjects for the message types (nine male, two female) and another group of 11 from the same pool of speakers served for the other two word lists. Each entry in the Recognition Accuracy column in Table 3 is based on a total of 1,100 entries of the word for types and fixes and 550 for digits; thus, each is considered quite reliable.

Most of the words in the three subvocabularies of this language were recognized with an accuracy of 98 percent or better. This figure does not include a rejection rate that also averaged about 1 percent (i.e., the utterance was not recognized as acceptably close to any of the reference images of the list at all). The error data are considered more critical, since the speaker's attention can be called easily to a "rejection" while misrecognition must be detected by the speaker himself.

To take one example of a rough comparison between spoken and key entry, consider the list of geographic fixes. Key entry of each requires striking three keys in an "artificial language." Thus, our 11 speakers made entry of each of 21 fixes 100 times or a total of 23,100 spoken entries. Overall accuracy of recognition was 99 percent

## TABLE 3

## VOICE ENTRY SUBVOCABULARIES

D-CONTROLLER MESSAGE TYPES

| WORD | KEY CODE | RECOGNITION ACCURACY* |
|---|---|---|
| AMEND | AM | 97.73 |
| CANCEL | CN | 94.36 |
| CORRECTION | CR | 99.73 |
| DEPARTURE | DM | 98.18 |
| DISCRETECODE | DQ | 99.91 |
| READOUT | FR | 99.91 |
| ACCEPTHANDOFF | AHO | 97.55 |
| HANDOFF | HO | 99.18 |
| DROPTRACK | DROP | 99.64 |
| PRINTSTRIP | SR | 98.82 |
| HOLD | HM | 99.82 |
| RELEASE | REL | 100.00 |
| REPORTALTITUDE | RA | 98.09 |
| WEATHER | WR | 99.36 |
| TRANSMIT | XM | 97.09 |
|  |  | (98.62 Overall) |

DIGITS (IDENTITIES, SECTORS, DATA)

| WORD | KEY CODE | RECOGNITION ACCURACY** |
|---|---|---|
| ZERO | 0 | 99.82 |
| ONE | 1 | 97.82 |
| TWO | 2 | 100.00 |
| THREE | 3 | 99.82 |
| FOUR | 4 | 99.09 |
| FIVE | 5 | 99.09 |
| SIX | 6 | 100.00 |
| SEVEN | 7 | 99.64 |
| EIGHT | 8 | 96.36 |
| NINE | 9 | 98.91 |
| ERASE | -- | 99.64 |
| BACKSPACE | -- | 100.00 |
|  |  | (99.18 Overall) |

*Each entry based on 1,100 spoken inputs
**Each entry based on 550 spoken inputs

## TABLE 3 (Continued)

FIX NAMES

| WORD | KEY CODE | RECOGNITION ACCURACY* |
|---|---|---|
| WILLIAMSPORT | IPT | 98.82 |
| SELLINGSGROVE | SEG | 98.91 |
| MILTON | MIP | 96.45 |
| HAZELTON | HZL | 97.45 |
| WILKESBARRE | AVP | 99.73 |
| EASTTEXAS | EXT | 99.91 |
| LAKEHENRY | LHY | 99.91 |
| TOBYHANNA | TSD | 99.45 |
| ALLENTOWN | ABE | 99.82 |
| STILLWATER | STW | 99.82 |
| BENTON | 7QB | 98.64 |
| SWEETVALLEY | 7EV | 99.91 |
| LOPEZ | 7LE | 99.82 |
| SNYDERS | 7YX | 99.73 |
| SLATINGTON | 7ZO | 99.36 |
| WHITEHAVEN | 9WT | 97.27 |
| RESORT | 9ZT | 99.27 |
| PENNWELL | 7PW | 99.73 |
| HUGUENOT | HUO | 98.45 |
| SOLBERG | SBJ | 99.09 |
| FREELAND | 7FE | 99.18 |
| | | (98.99 Overall) |

*Each entry based on 1,100 spoken inputs and approximately 1 percent of entries were rejected entirely. Key entry of these same characters would require 69,300 keystrikes with essentially no protection whatever from single-key errors, while each voice entry results in display of the whole three-character code for the fix which seems more susceptible of error detection than one or more misstruck keys. The time involved in the two entry methods seems indistinguishable. We did not collect accuracy data on key entry of fixes but this will be an integral part of later experimentation wherein voice versus keyboard entry of complete messages will be tested.

## Follow-on Reliability Studies

While the recognition accuracy data for the subvocabularies of this language were impressive overall, two major considerations impelled us to seek methods of improvement. In the first place, it must be remembered that the "user" here is the air traffic controller and the principal aim of voice data entry is reduction of distraction from his or her main concern, namely continuous observation and management of the dynamic four-dimensional traffic situation. It is thus essential that detection and correction of data entry errors be brought to some irreducible minimum. The second problem is that of individual differences in recognition accuracy from speaker to speaker. While precision and clarity of speech are of the essence in air traffic control, some controllers necessarily will speak with greater uniformity than others. Thus, while the overall recognition error rate for the message types subvocabulary was less than 1.5 percent, individual speaker error rates ranged from less than 0.1 percent to nearly 7 percent. With the "digits" subvocabulary, the overall average error was less than 1 percent while the range was from zero to 2.3 percent. Similar results were obtained for the subvocabulary of fix names.

It was decided, therefore, to investigate means of error reduction and/or error correction which might be applied to the basic VIP-100 recognition algorithm. We consulted with Dr. Breaux of the Naval Training Equipment Center regarding some of the recognition subroutines that he had developed for increasing recognition accuracy in his application in the ground controlled approach trainer. These, as well as a variation of the same general concept which was developed for us by Mr. Cox of Threshold Technology were experimentally tried with the non-radar controller data entry language with which we have been working. The net result, despite manipulation of the parameters of these routines, was either an increase in rejected inputs or an increase in the error rate or both. In retrospect this should not have been surprising, since the logic of these techniques is directed principally to the solution of the recognition problem where the input utterances are relatively long and largely identical with the exception of a single element. For example, the expressions "slightly (above/below) glidepath" can be differentiated with greater accuracy if both the reference and the input images are pared down to only those parts which are non-identical and a "second look" taken at the correspondences. This precise situation did not obtain in the word lists used here. The more common type of problem encountered was confusion of some of the pairs of words within a subvocabulary. The words "transmit" and "printstrip" in the message types list and the words "Williamsport" and "Resort" in the fix names list were among the frequent confusions. Oddly enough, even though the expression "nine" (instead of "nine_r_") was used in the digits word list, and nearly all errors involved the five/nine and nine/five confusions, a very high order of accuracy was obtained for both words.

In the course of trying out various alternative decision subroutines for error reduction and in re-examining our original detailed data we were struck by some interesting features of the word durations. For every utterance in the original tests we recorded the word numbers and correlations for the best and second-best matches and the duration (i.e., number of audio samples) of the input utterance. In the course of time normalization of utterances, we had been discarding this information after use. It was an interesting curiosity of our subvocabularies that some of the confusions that were common (such as Williamsport/Resort and fix/backspace/erase) were quite reliably distinguishable on the basis of utterance duration. In the course of investigating the utility of this phenomenon in turn (we started collecting utterance duration data during the "training" or reference array construction mode of operation) we further discovered that there were systematic differences in utterance duration during "training" as versus "recognition." The average duration of the utterance spoken repetitively during training frequently differed from the average duration of the same utterance spoken in a pseudorandom sequence. Since the durations differed under the two conditions, it was hypothesized that the correlations obtained in recognition would necessarily suffer.

The software was then modified in two ways. First, training was changed so that the speaker was presented with a pseudorandom prompting list. He or she did not simply repeat each word in the list in times in succession, but rather in times within the same list but seldom or never the same word twice in succession and in an unpredictable order. At the same time, the average duration of each word as well as the shortest and longest obtained during training were recorded and made a part of the reference information. The recognition decision algorithm was changed to make use of the duration data. The basic logic is as follows:

1. The input utterance is digitized, time normalized and its duration is noted.

2. The normalized feature array is compared with reference arrays for all words in the subvocabulary and the routine returns with the correlations for the best and second-best matches.

3. If the correlations differ by more than 40, the best match is selected as correct.

4. If the correlations differ by 40 or less, the input utterance duration is compared to the average (during training) duration for the first and second choice words unless the latter two durations themselves differ by less than 30 samples.

5. If the duration of the instant input is closer to the reference duration of the first-choice word, it is accepted as correct.

6. If the duration of the utterance is closer to that of the second-choice word, the utterance is rejected.

7. If the two reference durations differ by 30 samples or less, the test is not made and the first choice word is accepted as correct.

In addition to these changes in the training and recognition algorithms, we added a "tuneup" mode of operation to the basic program. In this mode of operation, the speaker puts on and adjusts the headset, adjusts the input volume setting and then starts reading the words in the particular subvocabulary. The recognition decision word is displayed on the Tektronix terminal CRT and just below it, the duration in samples of the utterance just made and the average duration of the first-choice (or recognition decision) word. If the two durations are not reasonably close (i.e., differ by more than 10 or 15 samples) for several of the words, even when repeated several times, then the headset placement and volume setting are rechecked. This "tuneup" mode is also useful for checking the effects of a cold or other speech altering event and the need for "retraining" specific words.

## Follow-on Results

Having made new training data by the pseudorandom repetition method, two of the "better" (i.e., higher overall recognition accuracy) and two of the "poorer" speakers were retested on the three subvocabularies previously used. With only one exception (fix names for one of the "better" subjects) the difference between the average duration of utterance in the training or reference data and the average duration of the same utterances under recognition conditions decreased substantially. With another similar exception, the average correlations of input utterances increased. That is to say, the quality of the matches between the inputs and their reference images, on the whole, improved. As might be expected, overall errors of recognition were reduced. The percentage error across all speakers and all three word lists went from 1.0 down to .35 percent. The percentage of rejects, somewhat surprisingly, went from 1.3 down to .8 percent. This last is surprising because it was expected that the use of duration information in the recognition decision logic would tend to increase the reject rate by rejecting some doubtful, atypical but correctly recognized (on the basis fo correlation alone) spoken inputs. This was a trade we were willing to make, namely, the exchange of rejects for errors. The "cure" for a rejected entry is simple: Say it again. The cure for an error is another story entirely.. Thus it would seem that the modified training routine alone solved most of the problem we sought to solve. In addition to this effect, the duration test in the decision logic only slightly increased the reject rate for two of the speakers on the list of fix names while the error rate for both was reduced to zero. Indications are, overall, that use of this additional information will convert a portion of the potential errors to rejects for some talkers.

Recognition reliability or error rate improved for both the "poorer" and the "better" talkers on all three subvocabularies with only two exceptions wherein it simply remained the same. In one of these two cases the error rate was zero under the original test conditions and, obviously, could not have been improved in any event. The improvements for the "poorer" talkers were not uniformly dramatic but they were very impressive in most cases.

It must be admitted that in the follow-on studies reported here we were proceeding on a "pilot-study" or "cut-and-try" basis until the very end. Thus, the final results noted just above are accounted for by a combination of variables. The training procedure was changed, the "tune-up" feature was added and the decision logic was modified. In addition, there may have been some unknown quantity of "Hawthorne Effect" upon the "poorer" talkers who worked closely with the experimentors through the cut-and-try phase of experimentation. The "acid test" of the objective changes should properly be made with a new sample of subjects but it does not seem likely at the present time that we will be given the time and resources necessary to accomplish this. On the whole, however, we feel that we have substantially realized our goal which was reduction of recognition error as close to the vanishing point as possible given the technology at hand. We believe that perhaps three to five errors of recognition in a thousand entries is a tolerable level within which to pursue further the applications which we have in mind. We fully expect that this error level will increase to some degree under conditions of lengthy message assembly as distinguished from subvocabulary testing. It remains to be seen how much it increases and what the subsequent ramifications (in user acceptability, for example) of such an increase may be.

## Other Findings

A great deal of ancillary but, from the standpoint of our potential applications, relevant and important information was also obtained. Our data from the initial reliability testing were studied for information on questions about speaker learning or familiarization effects, effects of such factors as colds and allergies on the recognizability of speech, effects of different types of microphones, and of precision placement of microphones.

In the matter of user familiarization and training, several important observations were made. During the test series for each speaker with each word list, recognition accuracy and "rejection" data were processed not less often than after every second session. As a rule, in the event that any individual word was either erroneously recognized two or more times or rejected as unrecognizable two or more times, a new set of "training" data was made for that word (and, in the case of errors, for the word with which it was confused if the confusion was consistently between the same two words). Thus, as recognition testing proceeded, the quality of the reference images or "training data" for some of the words in each list for some of the speakers was progressively refined. This does not mean that a great deal of retraining was done. A number of the speakers never needed to "retrain" any of the words in any of the lists at all. On the average, each speaker needed to retrain one word one time for the list of fixes, for example. Some speakers needed to retrain more words than others and some of the words and word pairs were more troublesome than others. See, for example, the fixes Milton and Benton in the list of fix-names. Attempts by some speakers to adopt an extraordinary (for them) pronunciation or emphasis in an attempt to improve recognition of a word were disastrous. Habitual or "natural" expression of the utterances is vital to accuracy of recognition. The modified training routine and our version of a "second look" in the decision logic (plus the long familiarity of the subject speakers by the time these were tested) reduced the retaining requirements to nearly zero.

Colds and allergies which affect the characteristics of speech were found to deteriorate recognition quality. However, for two of three speakers who among them contracted three head colds and one allergy during the test series, no serious problems were encountered. For these two speakers, it was necessary to retrain only a few of the words in the list to recover the near-perfect recognition previously found. One speaker, indeed, contracted a second cold after several weeks. It was only necessary to read in to the system the training data modified for the first cold in order to achieve the same recognition quality as produced by the "normal speech" training data. The third speaker, however, despite major efforts at retraining specific words was unable to regain a high recognition accuracy while the cold persisted. It should be noted that the overall data for recognition of message-type entries which has already been discussed includes the error data from this speaker which accounts for approximately half the total errors encountered with this particular subvocabulary. When this speaker was not suffering from a serious cold, his results were quite comparable to those of other speakers.

Retests were also run with most of the original twelve speakers using the last (and best) set of training, or reference, data recorded during the initial reliability testing phase. Retests were made after approximately 3 months and again after approximately 6 months following the last of the original test series. Both accuracy and reject results were almost identical to those found in the initial test series.

Finally, microphone quality and placement were found to be factors of influence. While fully systematic testing of these variables was not conducted, three different (but all "noise canceling") microphone types with four different mountings (one hand-held, three headset or headband) were employed at various times. The hand-held microphone was used by three of the speakers during the testing of the 15-word message-type list and accounts, in part, for the slightly lower overall accuracy rate found for that list than for the others. Careless, inconsistent, or unusual placement of microphones (e.g., at or below chin height, more than an inch from the corner of the mouth in the horizontal plane) immediately appears in a high reject rate because of loss of signal strength and can quickly be corrected by the user. Throat-type microphones were not tested but might be worthy of trial. The microphone used by all but one subject for the "digits" subvocabulary is directly substitutable in existing air traffic control operations for the carbon-type microphones required by the communications systems employed today. This microphone produced excellent results. Some further testing using actual carbon microphones belonging to field operations is planned.

## FY-78 PLANS

Over the next year we plan to complete at least some of the original overall experimental application plan which includes:

1. Experimental data collection in a series of "keyboard vs. voice entry" experiments. One of these is expected to be a laboratory, baseline establishment type of effort. A number of operators will simply make entry of a large number of traffic control computer input messages by both methods. This will provide a solid basis for: (a) Assessing the absolute and comparative reliability and efficiency (as

well as "user acceptability") of word recognition technology in this type of application, and (b) Assessing the subsequent effects of task-induced stress, the mixture of air-ground-air voice communication with ground-ground (controller/computer) communication by voice and other factors as yet unpredictable.

2. Addition of voice-feedback or audio verification to the system. Preliminary results indicate the possibility of no real gain in speed of entry but hard figures on accuracy of key vs. voice are not yet in hand. The principal gain we envision is reduction of distraction, a significant safety factor. Audio message verification may be an essential element of this last.

3. Field testing. Everything else being equal, a miniaturized (micro-computer) version of the "final" design will be (we hope) brought to a number of operational control facilities for field operator evaluation.

In subsequent years we will possibly experiment with language translation. The audio feedback technique we plan to use (which has, by the way, already been built in-house and is being tested) is based on digitization of real speech on a high sampling rate, limited feature count basis, similar to that used by long line telephone systems. We store these digitized images (practical at present for only a very limited language), concatenate and reconstitute them. The voice output quality is excellent and there is no problem of synthesis, especially of multiple natural languages. There is no raw synthesis, in fact, merely re-conversion from digital to analogue.

We have most of the software and nearly all of the hardware necessary to undertake these activities. The principal deficiency is people just now--the time of the principal investigator and the availability of subject/talkers.

## POST FY-77 TECHNOLOGY REQUIREMENTS

For the applications which we envision (and probably for many others) the following seem to be the "breakthroughs" needed to absolutely assure the future of voice technology in real-time, interactive command and control:

1. In the "word recognition" area (as distinguished from speech understanding), better word boundary detection. Maybe this really means limited SUS--for three and four digit numbers, for example. Many of our operational key-languages consist largely of numerical entries. Speed is important, here of course. None of our applications can wait seconds, much less minutes, for rendition of composite or even continuously uttered two to four digit expressions.

2. "Better" microphones or "less sensitive" (but equally accurate) digitization or both. What is meant here is a solution to the problems of speech-type noise, the necessity of precision microphone placement, the necessity of "calibration" of digitizers to microphones, some of the as yet unknown problems of speaker stress and similar accuracy-lowering factors. The solutions here might lie solely or principally in software, though possibly "adaptive" or "intelligent" hardware or some combination.

3. Continued improvement of the "many speaker" capabilities of SUS's. While the applications we envision can get by with pretrained systems (and, indeed, necessitate the precision and speed of the single-speaker, word-recognition technology as versus those of the present state of SUS technology) since we always know who the operator is going to be, these improvements would certainly be in the "nice to have" class. For some civil aviation applications, this characteristic is virtually an essential requirement. Applications here include general aviation pilot briefing and audio flight-plan entry.

## BIOGRAPHICAL SKETCH

### Donald W. Connolly

Engineering Research Psychologist
National Aviation Facilities Experimental Center
Federal Aviation Administration

1967 - Present: Primary responsibilities in human factor engineering of air traffic control ground environment - ATC facilities. Research, development, test and evaluation in man/machine interface technology - displays, controls, workplaces and operating procedures.

1952 - 1967: Previous experience in human engineering of Air Force, Army and Navy command and control systems and in basic and applied research in human factors with New York University College of Engineering and as a civilian with the U. S. Air Force.

Education: B. S. '50, M.A. '52, Ph.D. '57, Fordham College/University, New York, N.Y., Experimental Psychology.

C-3

## DISCUSSION

## Donald W. Connolly

Q: Mike Curran: You envisioned a success rate, 995 or 997 out of a thousand tries. That's pretty good.

A. I know that as we add factors of complication (stress factors), in other words, full message composition vs. just plain vocabulary testing, this rate will deteriorate. There is the question, for instance, of the task induced stress of actual control of traffic in a tense situation. It may or may not be a problem. But it probably will be a problem. One of the things I have discovered myself is that if you sit there and talk for four hours something happens to your voice. You may have to compensate for that too in some way.

Q: Mike Curran: Don, I didn't ask you the question yet. I'm giving you that success rate of 999 out of a thousand presuming we can ever get there. In your application do you still see a need for verification before the entry? Do you see the approach as a non-verified entry?

A: I think probably not, in the practical day-to-day operational sense. For instance, data entry, an error in the entry of data into the flight plan file is not fatal. It's lots of things, but it is not ordinarily fatal. Errors in communication between the controller and the pilot can be very serious. On the other hand, I see some kind of verification at least nice to have. Something which perhaps could be turned on and off, I don't know. At least for the beginning, absolutely essential. We have a redundancy in everything we do now. We still have the paper flight progress strips that they used in 1939, all racked up, just in case.

Q: Don Hansen, ONR: Given the verification then, and I guess I have to give you a 103 word vocabulary, what do you see as the minimum accuracy that you would accept with a "go" system to go to your agency and say this is it? You've got 997.

A: One of the things which I will find out, with any luck at all, in the next three or four months will be a direct comparison with the key entry system. I do know this, the language that they have to talk with their fingers is sufficiently artificial that a significant fraction of messages which are attempted to be entered with the fingers are flat out rejected because the format is wrong. Some kinds of errors are detectable. If you're controlling a high altitude sector and you try to put in an assigned altitude of 5,000, it will light up and say "tilt". On the other hand, if you intend (in a high altitude sector) to put in 37,500 and you put in 37,000, this is undetectable. You've got to detect that yourself. So the answer to your question is I honestly don't know. We worked with a guy some years ago who did an ops-analysis type study on the possibility of a mid-air in certain circumstances. He came back with possibly one in a billion operations, or something like that. And his boss said that we will never be able to publish that. The Answer is that we don't have any. We don't intend to ever have any and, in point of fact, that is the truth.

Q: Michael Nye: You quoted a rate, a data entry rate, a manual entry rate where you suggested that in the speech recognition test that you ran, you simulated 70,000 key strokes.

A: That was simply a total aggregate, a key-by-key comparison between 23,000 voice entries which were translated into the equivalent of 69,000 keystrikes.

Q: Michael Nye: I understand. I guess my question is one of two parts. The first part is that we make a big to do about the accuracy of the speech recognition device but we don't talk about the accuracy of the human's ability to be able to enter 70,000 key strokes in the right kind of format.

A: You're absolutely right there. The only people who can do anything like that are professional keypunchers who do nothing but punch keys. Air traffic controllers are never going to be in that business while I'm alive.

Q: Michael Nye: O.K. It's then safe to say that the accuracy, no matter what the accuracy is, as long as its above 98%, is probably more accurate than the manual method.

A: I intend to find that out in the next couple of months.

Q: Michael Nye: The other question was: you made the statement that it does not appear that speech recognition offers any advantage in terms of input speed or throughput and I challenge that and I'm curious.

A: Well, at present if you take a hunt and peck operation where the shifts of attention are involved as well as remembering and constructing this artificial language, we're talking about three or four keys a second. In other words, the coded key equivalent of the message, we're talking about three or four keys a second, that's about what you get out of voice entry on all the preliminary data I have now. Speed is not so much the name of the game. I can conceive of applications where the speed would be much greater. That just happened to be an observation in passing.

Q: Michael Nye: O.K. I guess the point of view that I had was that you're looking at an isolated application where the speech recognizer is reduced to working in an environment that simulates a keyboard entry routine whereas the real benefit of this kind of technology is in applications where you eliminate the manual method of data capture and you use a voice method. In essence, instead of saying a series of code numbers or code words you actually say the phrase, the phrase is entered in a split second whereas to reduce that to a manual method, maybe four or five key strokes, and in that application, there is a tremendous increase in throughput. Is that true?

A: As I said in the beginning, at least initially I'm working with "unnatural" languages, if you'll pardon the expression.

Q: Don Hanson, ONR: Your vocabulary of 103 may be picked to give a good result. What if you elected to increase the vocabulary? What if you jumped to 1,000, what would you expect your accuracy to be?

A: They probably wouldn't ever do that in the whole class of applications that I'm talking about. The total language of air traffic control, the total human language, is not over 300 or 400 utterances. That's one point. Secondly, as long as your subsets are small enough (and other people have alluded to this) accuracy need not suffer. If you get a subset over 40 or 50 elements and especially, for instance, in this case of fixes, place names, and this sort of thing in the real world you're going to run into things which are sound alikes and you're going to be stuck with them. You're going to get some reduction in accuracy, no doubt about that. Big enough subsets big enough possible set of confusions, errors go up.

Q: Don Hanson, ONR: No, I was thinking of your experience. What's a group figure? 80%, 90.

A: Wouldn't even guess it at a thousand words. No. Not from my experience.

Q: Danny Cohen, Information Sciences Institute, USC: I really appreciate the comment about phrase recognition and I think that the right way to go is by recognizing phrases and being more one mutual language. Does anyone have statistics for recognition of phrases? Is the number still 99.3 or is it more like 60 or 40?

A: Well, I'm working strictly in isolated utterance recognition. Now an utterance can be a whole phrase. When we get into speech understanding where there is some interpretive and analytic work involved in the understanding of what the constituents and the sense of a phrase are you're getting out of my field and I don't know the answer.

Q: <u>Leon Ferber</u>: I wanted to ask you whether you did any studies which probably has nothing to do with speech recognition but there is this phenomenon that either you look at the display or you look at a display and you talk you're just as effective as if you just look or just talk. It's this effect of walking and chewing gum.

A: Yes. We have a little data on this. We've done some work for instance with head mounted eye cameras and this sort of thing and many times the human operator <u>seems</u> to be functioning in parallel fashion or simultaneous fashion but he is really switching back and forth between a couple of sequential operations. I think the truly vital thing is not to lose the picture and if you must look away, you will lose the picture and you will lose some of the important parts of the picture and these are the things that are now possible through computers such as the conflict avoidance alerts. Its one thing to hear a bell and look back and see what's blinking and then you try to figure what the heck it is, another thing when you've got your eyes on that display all the time.

Q: <u>Bob Fleming, Naval Ocean Systems Center</u>: I was wondering if this was introduced to air traffic controllers. Have you given any thought to the mechanics of switching in and out of when he is talking to another aircraft versus talking to the system itself?

A: That is one of the very definite operational problems that we're going to have to face. I see it as a detent microphone switch or something of that sort at least for a starter. One of the things I see in the end going toward the sublime, I see much of what a controller has got to tell the computer with his fingers he has already told a pilot with his mouth. Now if we can just pick out what the computer needs to know from what he told the pilot we have it made. Thank you very much.

D10

[redacted]

N93-72622

# MILITARY APPLICATIONS OF AUTOMATIC SPEECH RECOGNITION & FUTURE REQUIREMENTS

50-32
176346

DR. BRUNO BEEK
EDWARD J. CUPPLES

ROME AIR DEVELOPMENT CENTER
GRIFFIS AFB, ROME, NEW YORK

The objective of this paper is to provide an updated summary of the state-of-the-art of Automatic Speech Recognition (ASR) and its relevance to military applications. This paper follows the format of the recent IEEE paper on "An Assessment of the Technology of Automatic Speech Recognition for Military Applications (I)". Until recently, speech recognition has had its widest application in the development of vocoders for narrowband speech communications. Presently, development in ASR has been accelerated for military tasks of Command and Control, Secured Voice Systems, Surveillance of Communication Channels and others. Research in voice control technology and Message Monitoring Systems and Automatic Speaker Verification Systems are of special interest. Much of the emphasis of today's military supported research is to reduce to practice the current state of knowledge of ASR as well as directing research in such a way as to have future military relevance.

Already in use are voice data entry systems for a number of interactive command and control functions. These applications are limited to a small vocabulary, speaker dependent, isolated word recognition system. These highly reliable systems allow for hands-free source data entry of the digits and a limited set of control words and phrases. As such, the voice data entry system eliminates manual transcription and keying operations in hand/eyes busy mode of operation.

A large number of potential systems for military applications are now under development. These include:

1. Digital Narrowband Communication Systems: A large research and development program has yielded a number of digital communication systems mainly using linear predictive encoding analysis. These various systems have been analyzed and equipments are now being developed for operational tests. These efforts have resulted in the need for the fabrication of a low cost speech input/output front end. Major research in this area is underway.

2. Automatic Speech Verification: An advanced development model has been fabricated for secure access control applications and has been tested under laboratory and operational conditions. These tests demonstrated that speaker verification is a viable technique.

[redacted]

3. On-Line Cartographic Processing System: Studies are under way to use speech recognition and voice response techniques with cartographic point and trace processing systems. Equipments have been fabricated and tested under operational conditions. Other research and development programs are now undergoing test and evaluation.

4. Word Recognition for Militarized Tactical Data Systems: Word recognition, speaker verification and voice response will be used for message entry to a tactical data system. Equipment has been fabricated and is being or has been tested.

5. Voice Recognition and Synthesis for Aircraft Cockpit: Existing word recognition systems are being tested and evaluated under simulated cockpit environments.

These system studies and implementation are in various stages of investigation and/or development and represent a practical utilization of the emerging speech recognition technology.

Numerous independent and integrated efforts have been undertaken to further develop the speech generation, reception and reproduction phenomena for specific military and civilian applications. This has increased the interaction among scientists in fields of acoustic phonetics, linguistics, signal processing, computer science, etc. Fortunate or unfortunate, the major emphasis on present day funding is dealing with the practical application with existing word recognition systems with only minimal consideration to some major improvements which are necessary before these techniques can be used in ever increasing military systems.

1. Automatic Message Monitoring: As the technology advances, the utilization of message monitoring functions shall find increased relevance for command and control functions.

a. Keyword Classification: The goal of keyword recognition/classification is to recognize a keyword or a set of keywords embedded in narrow-bandwidth conversational speech as expected from a radio link. The reconnaissance of large amounts of speech information requires a need for economical data editing and scanning. An automatic method of detecting and classifying keywords would perform this function. The difficulties in providing this speech processing function are that the speaker is unknown, the speech is continuous and the speech signal is often degraded by noise and distortion. Problems also exist due to coarticulation and context since the acoustic representation can be affected significantly by the acoustic environment of the keyword.

The technique which seems to have had the greatest amount of success is to recognize acoustic events simultaneously in free running speech. A sequential logic, made up of acoustic events, can be designed for a keyword. This approach postulates that keyword detection will take place when the needed sequence of acoustic events occur. Results to date indicate that detection rates are approximately 85% with 10-15 false alarms per hour.

b. Language Identification: Numerous techniques have been applied to Language Identification due to a fluctuating military interest in this problem. The most general approach has been, as in speaker identification, to use pitch and spectral features. Recently, emphasis has been placed on the recognition of acoustic events within the speech signal which are reliable, stable, and speaker independent. This work, and the work of devising methods of converting the acoustic signal into a chain of linguistic elements, offers considerable promise for a solution of this problem.

Experiments have been performed with perfect chains of this type, formed from a phonetician's hand transcription of speech in various languages. These experiments show that the statistics of these chains, especially higher-order statistics of digraphs and trigraphs, are a very powerful means of discriminating between languages given one to two minutes of speech.

Experiments were also performed on five languages involving over a hundred different speakers using acoustic events extracted from continuous speech. The data on each speaker was collected over a period of time to determine the effects of speaker variabilities on recognition performance. Fairly encouraging language recognition scores for all five of the languages has been attained. Continued research and development is necessary in order to implement an on-line, real-time operational system.

2. Command and Control: Command and control, in the context of this paper, means voice communications with machines. These include:

Limited Word Sets

Connected Word Recognition (Limited set of words)

Continuous Speech Recognition/Understanding

a. Isolated Word Recognition: Speaker dependent isolated word recognition for all intents and purposes has been solved under laboratory conditions, but engineering problems still exist. Even with these problems a number of commercial companies in the United States and England are marketing isolated word recognizers. Some voice date entry systems are already operating in a variety of industrial applications. These include:

1) Automatic Sorting Systems, distribution of parcels, containers and baggage.

2) Voice Programming for Machine Tools

3) Inspection System, e.g., measurement of TV face plates

4) Automobile quality control inspection

5) Various military applications.

Generally, two types of problems exist, those dealing with the speech processing/recognition process and those related to the incorporation of the word recognition device into an operational system. The first set of problems is:

1) Development of low-cost, light weight human acceptable, noise cancelling microphone.

2) Training various individuals to use the word recognizer. This includes minimizing the number of repetitions per word. This problem becomes more severe as the vocabulary increases.

3) Development of an adaptive technique to upgrade the word patterns as a function of time.

4) Decrease the error rate. Even a 1-3% error rate can cause large time delays in data input.

5) Others

System problems include the development of a gracefully interactive system that an individual can feel comfortable with. This includes:

1) A learning procedure. Teaching the individual to use the system in such a way that he uses it in a normal everyday manner.

2) Develop error correcting methods to limit backup time.

3) Development of optimum system feedback (visual, audio or both).

4) Others

b. Multi-Speaker, Isolated Words: Emphasis is being placed today on the development of an isolated word recognition speaker independent system which operates over a good quality telephone network (bandwidth 300-3500 Hz). Experiments have shown that a relationship exists between word size and speaker dependence as a function of recognition accuracy. Although most multi-speaker systems tend to be insensitive variations in the rate of speech, problems still exist due to variations in a talker's speech characteristic and talker dependent traits. Recent experiments have shown a multi-speaker recognition system to be highly accurate in recognizing a small set of words (10-20) even in the context of connected speech. However, a number of constraints had to be introduced. These include:

1) Knowledge of the number of words in a string

2) Fixed number of digits with error correcting codes

3) Requirement of a short learning phrase.

c. Continuous Speech Recognition (Speech Understanding): The Department of Advance Research Projects Agency has recenlty completed a program to recognize sentences that can be produced from a known vocabulary (Lexicon), with known subject matter (Semantics), using known grammatical rules (Syntax). The results of this program are now being published. In addition, a study is underway to determine the major scientific contributions of the work. The ARPA project considered the task domains as data retrieval.

3. Speech Enhancement: Programs are underway to enhance the intelligibility of speech signals transmitted over a low-quality communication channel. Techniques under investigation offer potential for the development of an automatic system for attenuating non-speech signals which accompany speech and interfere with and occasionally obscure the information bearing parameters of speech.

Two techniques were developed for enhancing the S/N ratio of speech received over a noisy channel. The first method is intended for use when the noise is wideband and random. It is similar to homomorphic filtering, however, the spectrum rooted (rather than logged) before being transformed. Limited test results showed the effective S/N ratio increased somewhat without seriously distorting the character of the speech.

The second enhancement technique is useful when the interference consists of tones or can be decomposed into tones. It consists simply of transforming the speech-plus-noise to the spectrum domain, detecting and attentuating the tones, and retransforming the enhanced spectrum to the time domain. By use of this method, speech signals that were barely detectable at S/N ratios below -26 dB were made fully intelligible.

The enhancement processing in its present configuration requires extensive hardware and software for real-time operation. The reason for this complex system is to transform the noisy speech signal in such a way as to easily remove the noise portion without appreciably affecting the voice signal.

Systems of this type are in various stages of exploratory research and advanced development. In the near future we shall see the operational implementation of these systems.

4. Security Applications: Automatic speech recognition may soon be extensively used in the area of security. First and foremost in terms of military applications is the problem of automatic speaker verification/identification. This security problem and others are in different stages of solution.

a. Automatic Speaker Verification (ASV): Speaker verification means the verification or rejection of an individual based on his speech patterns. In general, each individual known to the ASV System has on file a number of samples of his speech. When he wishes to be verified he must first identify himself to the ASV system via a badge reader, keyboard, magnetic card, etc. Once he has identified himself, the ASV system requests that he speak a number of sentences, phrases or words. After the individual complies, the ASV system analyzes the incoming data, compares it with its reference file and makes a decision either to accept or reject the individual or request additional data. ASV technology has a number of advantages not always available in other speech recognition technologies. Namely, the speaker is cooperative and is attempting to gain access to some function and hence will be on his best behavior. The speech data spoken by the individual is known to the ASV system; sentences and words are chosen to provide the greatest amount of discrimination. The acoustic environment can be either controlled (good signal-to-noise ratio) or noise cancelling microphones can be used. Analysis of an individual's reference speech data may lead to extraction of customized features for that individual to maximize speaker discrimination. In the operational mode, where the individual tests the system, the ASV can, by analyzing the speech and finding it deficient (not loud enough, garbled, etc.), can request individuals to repeat. Further, the communication channel can easily be made identical for both reference and test.

A number of techniques are being investigated by Bell Telephone Laboratories, North American Rockwell and others. Perhaps the most successful is that of Texas Instruments (supported by the United States Air Force/RADC).

This technique was designed to have less than 1% rejection of true speaker, and less than 2% acceptance of impostors.

The ASV system has been tested under the conditions of mimicry, day-to-day speaker variability, colds, sinus congestion, and respiratory ailments. In general, the technology can handle most operational requirements and achieve low Type I errors (true speaker rejections) as well as Type II errors (impostor acceptance) with certain tradeoffs depending on the specific application. For example, Type II errors can be lowered at the expense of increasing Type I errors and vice versa. Both types of errors can be reduced at the expense of having the user utter more speech data.

The technology has been tested operationally by the Air Force Electronic Systems Division (ESD/BISS) and Mitre Corp using a large group of military personnel. In addition the system has been in day to day use at the Texas Instruments Corporate Information Center for three years and has achieved a Type I error rate of 0.5% and 1.6% Type II error rate.

This is a research area of interest to the military (e.g., accepting or rejecting front-line tactical reports), in industry (e.g., controlling access to restricted areas), and in commerce (e.g., controlling access to money or information). There is a stated military requirement (ROC) for a device to handle this problem under field conditions. The technology is presently being advanced via the development of systems which operate over phone lines and systems which require an individual to only speak a 6 digit code for identification and verification.

b. Speaker Identification (Recognition): This section describes several problems dealing with automatic speaker identification of an individual based on his voice characteristics. Speaker identification is discussed in the context that recognition decisions are rendered automatically from continuous speech uncontrolled as to context. Hence, the speaker's reference library and testing unknown speech samples can be generated independent of spoken text. This advantage is important when one is collecting and analyzing speech data from an uncooperative speaker. At the present time, applications of greatest interest are the use of the voice signal for personnel identification for access control, monitoring communications channels and computer security. Conditions that make this problem difficult in practice or real application are:

1) the communication system is of poor quality

2) speakers may be non-cooperative

3) recording and/or channel conditions are different for reference and test samples

4) deciding whether the speaker is a member of an original group of speakers

This kind of problem arises in the military when, for example, one tries to keep track of a unit that is communicating by radio. Presently, emphasis is being placed on analyzing an individual's speech signal when particular phonetic events occur. In this manner, the vocal tract transfer function can be determined by a detailed spectrum analysis. Combining a number of these measurements for a number of phonetic events can provide the data to identify a speaker.

A number of speech investigations have studied the use of linear least-square, inverse filter formulation to estimate the format trajectories of selected phonetic events.

## Summary and Conclusions

Although significant progress has been made in all areas of automatic speech processing in the past ten years, we still have a long way to go before they can be incorporated into the military inventory.

We have attempted to describe the various military problems, existing solutions and how they lead to present and future technology requirements. The utilization of these automatic speech processing systems have lead to new emphasis in technology development, namely, the requirements for a low cost speech processing front end system to be used for automatic word recognition, speaker identification and vocoder applications. Consideration is being given, by various systems choices, to the "best" methods of incorporating word recognition technology as a peripheral device in their system specifications. Numerous other requirements could be listed, however they deal mainly with the technology in its present and near future development.

We must be careful that we do not try to force-fit the present day technology into areas that really require advanced speech processing concepts. For example, although it is true that isolated word recognition can solve a wide range of present day problems, care must be taken to discern the limitations of these systems with respect to a truly correct speech application.

The present day focus of attention is in the area of applications rather than basic speech research. The speech pendulum again reached an extreme position, limiting the amount of speech research.

We reiterate that more basic research into speech perception, linguistics, acoustic-phonetics is necessary before a thoroughly adaptive error free speech processing system can be constructed.

## BIOGRAPHICAL SKETCH

### Dr. Bruno Beek

Dr. Bruno Beek was born in [redacted] on [redacted]. He received the B.A. degree in physics and mathematics, and the M.S. and Ph.D. degrees in physics from Syracuse University, Syracuse NY, in 1956 and 1970, respectively. In 1956 he joined the Rome Air Development Center, Griffiss Air Force Base, Rome, NY as Research Physicist, doing work in electronic warfare technology. In 1962 he began doing research in automatic speech processing techniques. He has also participated in contract programs related to speech research and has been involved in the Department of Defense long-range planning program. In 1976 he served as a Speech Specialist to the AC/243 (Panel III) Research Study Group 4 in automatic pattern recognition under the North Atlantic Treaty Organization (NATO). He has contributed a number of articles, been invited to participate in national and international conferences and seminars. Dr. Beek is a member of the IEEE S-ASSP Speech Processing Technical Committee.

(This page intentionally left blank)

# THE ARMY WORD RECOGNITION SYSTEM

DAVID R. HADDEN
DAVID HARATZ

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

## 1. INTRODUCTION

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

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

## II. PRESENT STATUS

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

[redacted]

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

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

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

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

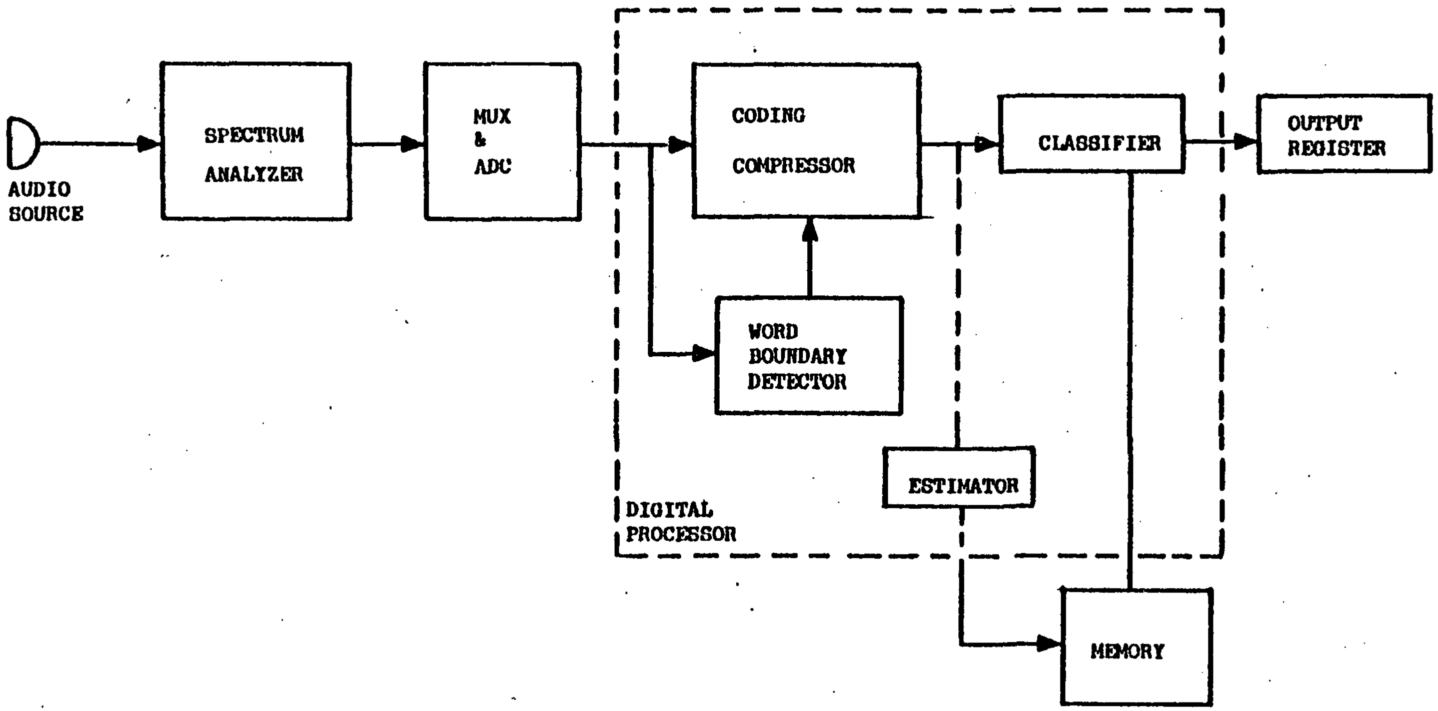


Figure 1. Word Recognizer in Army WRS

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

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

## A. Hardware Design

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

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

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

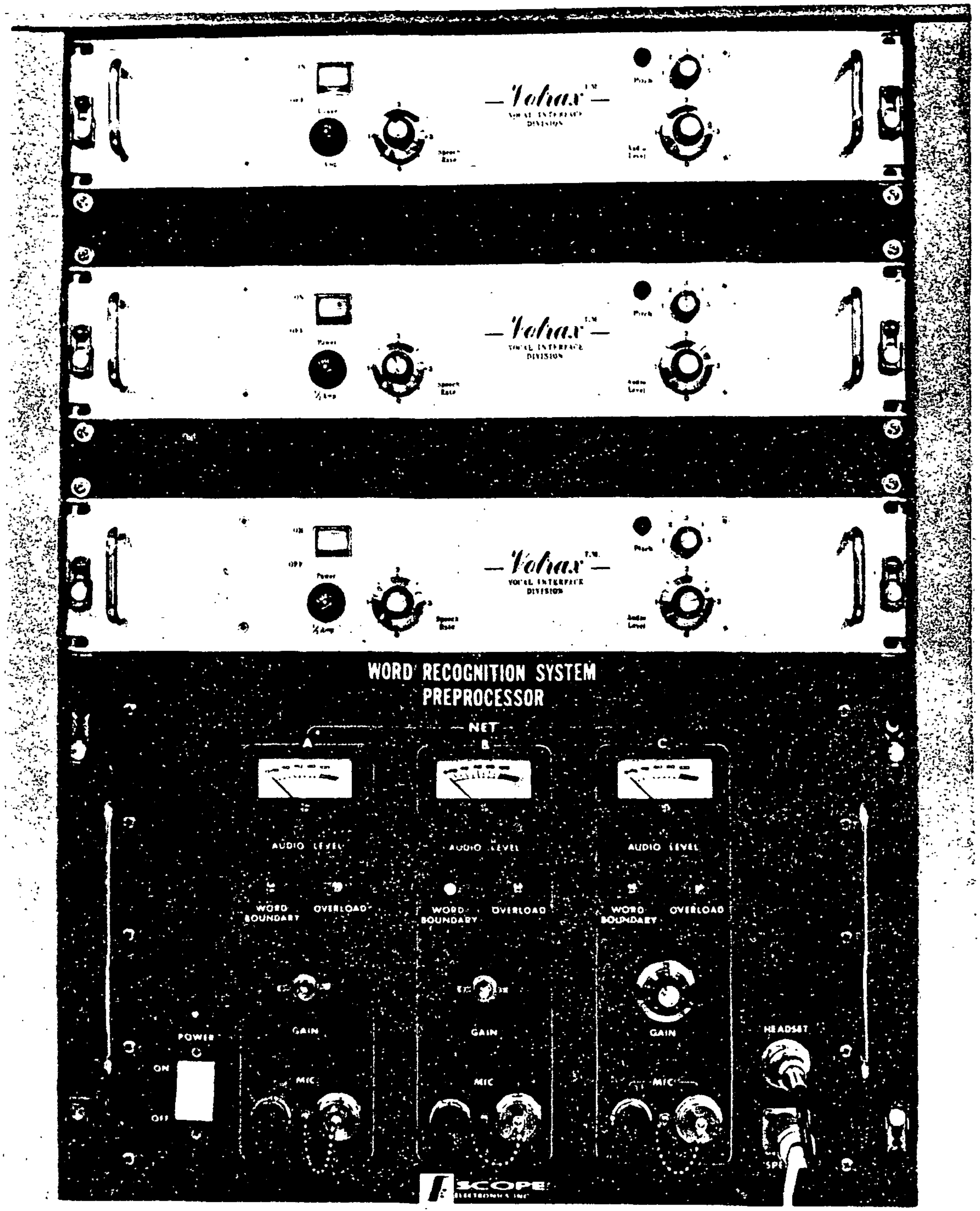


Figure 2. WRS Preprocessor Module and Votrax

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

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

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

## B. Software Design

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

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

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

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

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

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

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

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

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

### C. Results of WRS

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

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

## III. SUMMARY AND FUTURE DIRECTION

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

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

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

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

## ACKNOWLEDGMENTS

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

## BIOGRAPHICAL SKETCH

David R. Hadden, Jr.

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

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

(This page intentionally left blank)

D/12

N93-72624

[redacted]

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

52-32
176348

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.

D/13

N93-72625

[redacted]

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

S3-32
176349

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

[redacted]

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.

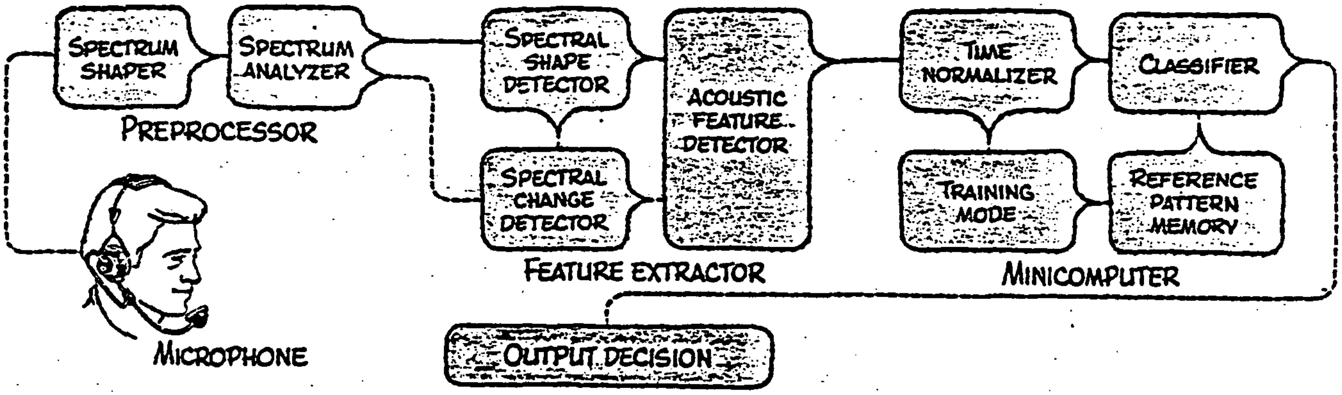


Figure 1. Block Diagram of Speech Recognition System

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.

```
    Spectral        Phoneme
    Features        Features
 1          **      *  **
 2**        *       *  **
 3**      * **      *  **
 4 **    *  ****        *
 5 **    ** * **           *
 6 ***   *  ****           *
 7 ***   ** ****     *      *   *
 8 ***   *  **** *   *          *
 9 **   **  * *  **  *       *  *
10 **   **  * *  **  *       *  *
11  **  **  ******              *
12  **   *  ******              *
13* **  **   ** **  * *      *  *
14*     *     * **    *  *   *
15*           *       *      *
16*                          *
```

Time

Reference feature matrix for the word "seven"

```
 1**        **      *  **
 2**        **      *  **
 3 *        **      *  **
 4**        ***     *  **
 5 **    ** * **           ** *
 6 ***   ** * **           **
 7 ***   ** ****     *     **   *
 8 ***   ** ****     *      *   *
 9 ***  **  **** *   *       *  *
10 ***  **  * * **   *       *  *
11  **  **  **** *   *          *
12  **  **  ******              *
13* **  **  ****** *  *  *   *  *
14*     *     * ** *  *  *   *
15*           *       *      *
16*                   *      *
```

Time

Feature matrix for "seven" to be compared with reference matrix.

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$^{R}$-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.

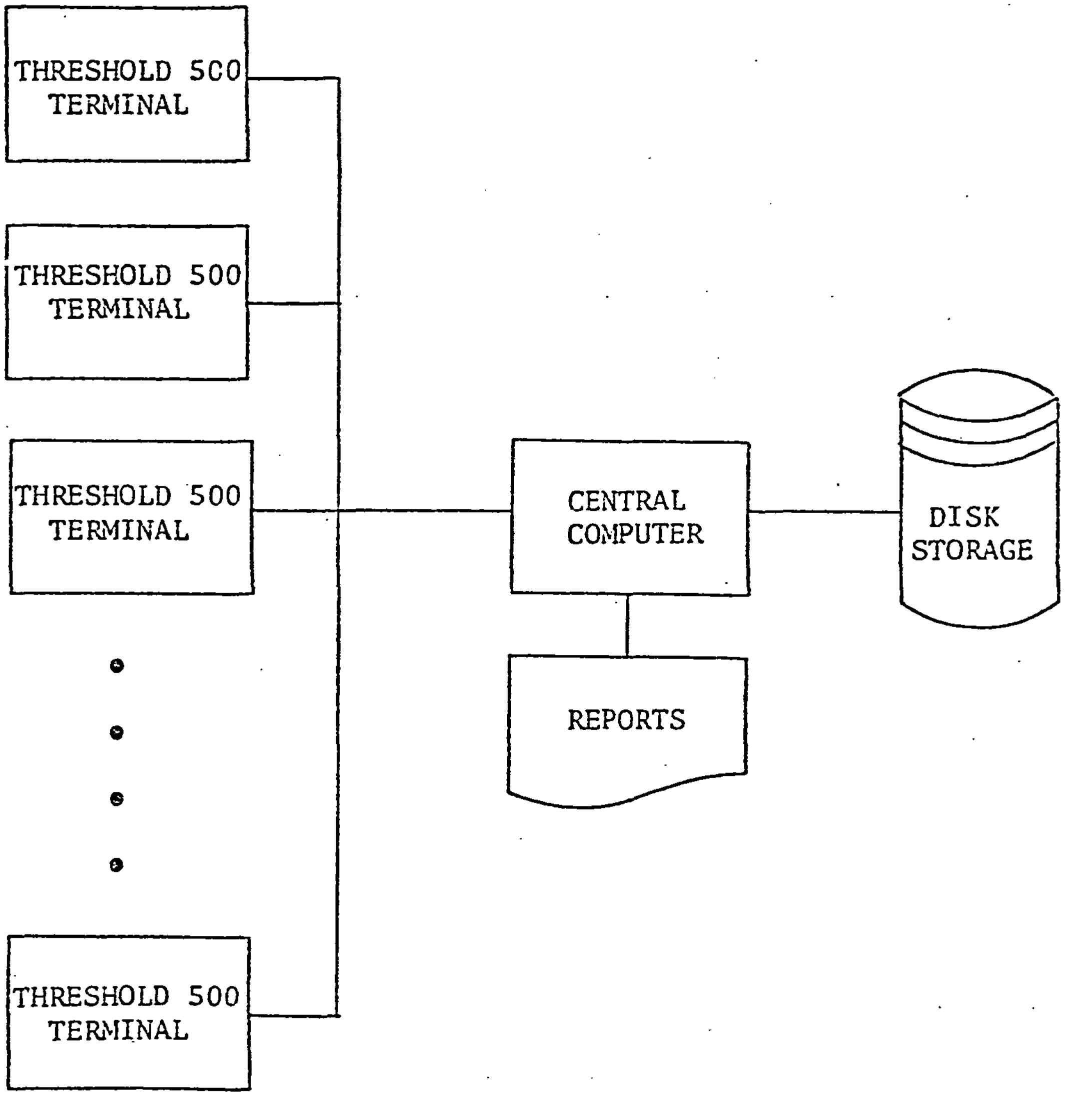


Figure 3. Threshold 500 Terminal System - Block Diagram

## 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

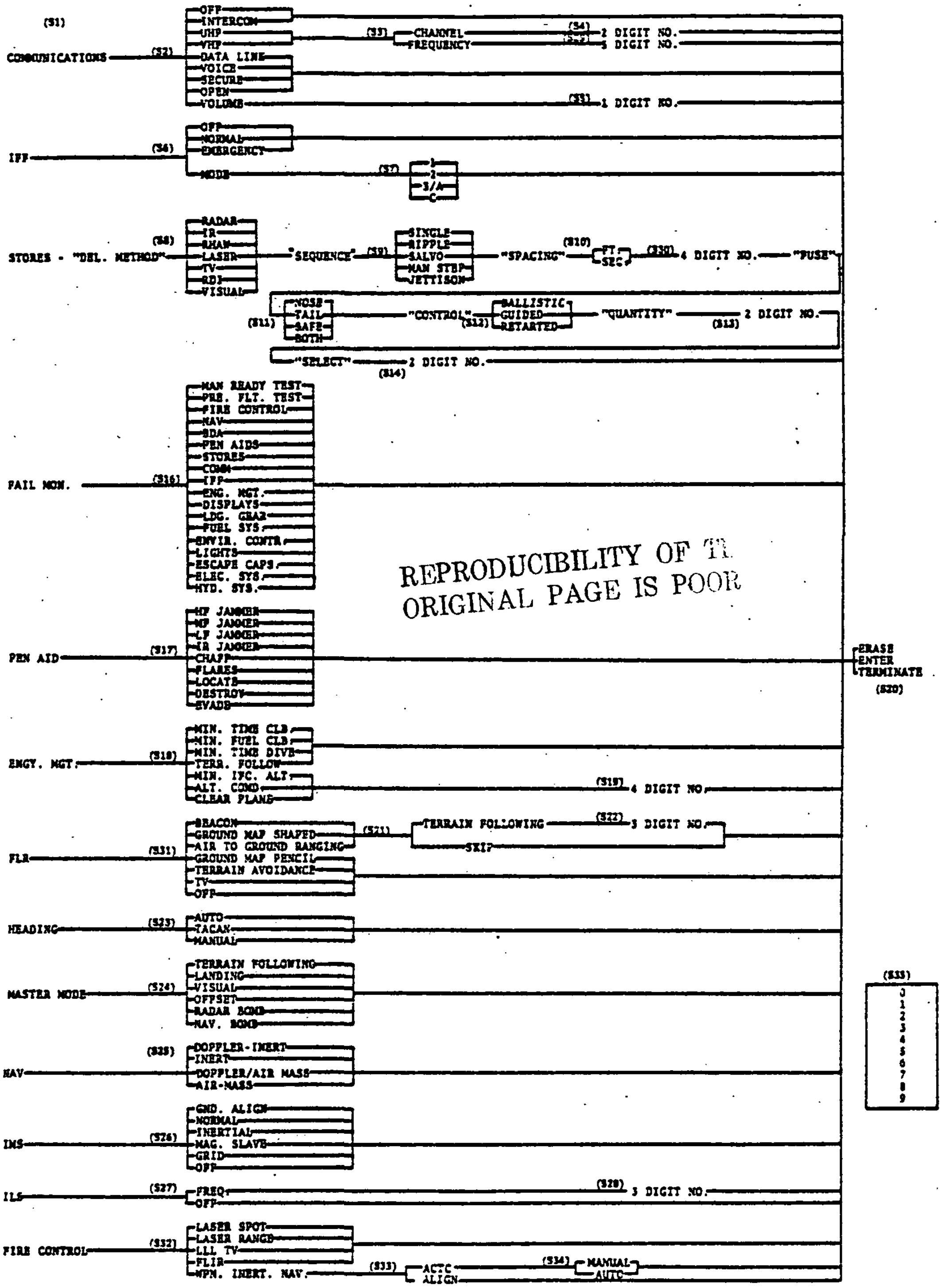


Figure 4. Voice Control System Command Structure

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 ) Invented hybrid logic suitable for real-time pattern recognition
1961 ) (analog-threshold logic).

1962 ) Demonstrated vowel recognition ) Isolated speech.
1963 ) ) Obtained highest
1964 Demonstrated consonant recognition ) 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 ) Developed and delivered miniaturized voice controller for astronaut.
1968 ) Developed NST to recognize digit strings for U.S. Post Office,

- operated in real-time in high noise environment
- for universal speech including many dialects, largest speaker population ever tested.
- 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 ) Introduced user-programmable voice data entry terminal (Threshold 600)
) which is CRT/Teletype compatible.
)
) 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.

(This page intentionally left blank)

omiT

## SESSION III

DR. EDWARD M. HUFF, CHAIRMAN

NASA AMES RESEARCH CENTER
MOFFETT FIELD, CALIFORNIA

(This page intentionally left blank)

[redacted] D14

N93-72626

## STATISTICAL ASSESSMENT OF SPEECH SYSTEM PERFORMANCE

S14-32

STEPHEN L. MOSHIER*

DIALOG SYSTEMS, INC.
BELMONT, MASSACHUSETTS

176350

### Introduction

Since many of the participants at this conference have the task of evaluating and comparing different speech recognition systems that have been tested under various conditions, the author has collected in Part I some useful statistical rules of thumb which can be employed to normalize disparate experimental test results. Several types of elementary statistical analysis are illustrated; the reader is encouraged to continue in this spirit to analyze other cases. The rules are not widely known, but seem to have good predictive power. All of the ones presented here are accompanied by supporting empirical evidence.

Part II, the advertising part, describes some of the accomplishments and planned development activity of Dialog Systems, Inc. in accordance with the Workshop specification. Dialog's sole business is speech recognition. The company has successful operational field experience with its first major product, a multi-channel talker-independent system for verbal inquiries via ordinary switched network telephone input. Dialog presently has 45 employees, including a competent support and field service staff.

### I. Methods for Normalization of Performance Test Results

Contrary to some beliefs, speech recognition systems obey the laws of nature. The small number of known quantitative rules are statistical in character, and they relate such variables as average recognition accuracy, vocabulary size, reject rate, false alarm rate, and the sizes of experimental training and test sets. The relations to be presented here seem to have good predictive power, and the author uses the illustrated analyses on a day-to-day basis to evaluate and compare different experimental results. It is very obvious that some such probabilistic rules must apply, though there are questions of detail and refinement of the statistical models to be resolved. It will be possible in the future to tie together many other experimental variables, but to do this it will be necessary for investigators to include more detailed experimental data in their reports and to test much larger populations than they have been accustomed to using, on the average.

*Mr. Moshier's paper was presented by Mr. Robert Osborn.

The ultimate capability of practical speech recognition systems has not been determined. The speaker verification system developed by Doddington's group, for example, probably does better than a human could do; the computer does not get tired, and can be programmed to notice identifying characteristics that people pay no attention to. In the various kinds of speech recognition, performance is limited by such things as insufficient data bases, mistakes in computer programs, and adherence to wrong theories; we are certainly quite far from any limits set by thermodynamics or information theory.

When various published data are normalized by means of the statistical rules to be described, it emerges that there has been essentially no fundamental progress in isolated word speech recognition since the first good techniques appeared in 1969-1972. On the other hand, there has been a great deal of progress in making the fundamental principles work in field applications, as well as in other areas.

## Vocabulary Size

Most reports on speech recognition give a figure for recognition rate and vocabulary size. The law of nature is that recognition rate decreases with increasing vocabulary size. It is quantified by a statistical rule of thumb as follows:

Given that the input speech is word (or phoneme, or sentence) $X_i$, suppose that the machine is characterized by the probability that it will correctly reject the possible wrong choices $x_j$, $j \neq i$:

$$\Pr\left\{\text{correctly reject } x_j | x_i\right\} = r_{ji}. \tag{1}$$

In the interest of deriving a simple formula, make the following assumptions: a) $r_{ji}$ is about the same for all pairs of vocabulary words, so that it can be replaced by a constant value r. (In practice this is usually true except for a small number of troublesome words having high confusion probability; but if the relative proportion of troublesome words is constant the formula remains true for an appropriate choice of r. Thus a less stringent assumption is sufficient, namely that the distribution of values $r_{ji}$ is dependent of vocabulary size.) b) The various correct rejection probabilities (1) are all statistically independent. This assumption permits getting the probability of joint events by multiplying. As in assumption a), it can be replaced by less strict conditions, but then the development becomes more obscure. It is <u>not</u> true for some types of joint events encountered in continuous speech recognition (see below). Let there be n words in the vocabulary; under conditions a) and b) the probability of correctly rejecting all of the wrong choices $x_j$, $j \neq i$ is

$$\Pr\left\{\text{correctly reject } x_{j_1} \text{ and } x_{j_2} \text{ and } \ldots x_{j_{n-1}} \mid x_i\right\} = \prod_{j \neq i} r_{ij} = r^{n-1}. \tag{2}$$

This is the correct recognition rate of the system. Curves for talker-dependent and talker-independent isolated word recognition are given in Figure 1. These curves were drawn in mid-1974; there does not seem to have been much change since then except that they are perhaps a little truer now than they were before.

False Alarms

A related statistical rule has to do with false alarm events in word spotting. In this task, the event of interest is the joint detection of several acoustic segments in the right sequence. The unconditional probability of a false alarm for any one of the segments is assumed to be small in a small time window and independent of time. The distribution of false alarms is therefore Poisson with some rate function $\lambda$. Given the acoustic event $x_1$, certain following events are more likely to occur than others, on the average. Thus it is not possible to get the probability of a joint event by multiplying the individual event probabilities. A joint false alarm can be modeled quite closely, however, as a sort of Markov chain. It is assumed that if the first target segment $x_1$ is detected, the unconditional Poisson rate function for subsequent detection of the second event must be multiplied by some value $\alpha$.

The Poisson law implies that the unconditional probability of not detecting the $i$th event is $e^{-\lambda i}$. If the conditional detection probability depends on the immediately preceding event but no earlier ones, the probability of the first two events jointly is

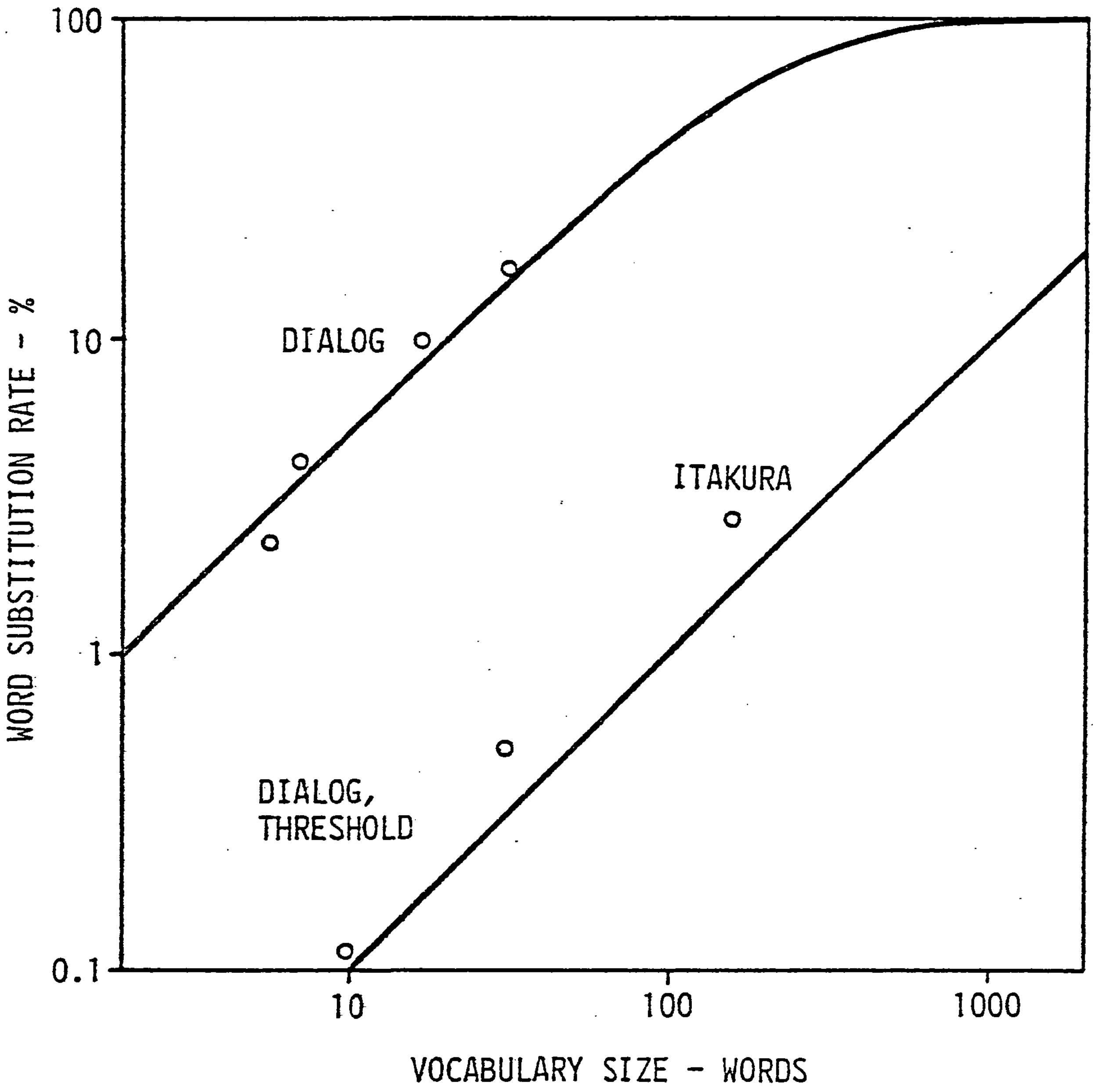
$$\Pr\left\{x_1 \text{ and } x_2\right\} = \Pr\left\{x_2 \mid x_1\right\} \Pr \ x_1 = (1-e^{-\alpha\lambda 2})\ (1-e^{-\lambda 1}).$$

The joint detection probability for n events is then

$$\Pr\left\{x_1 \text{ and } x_2 \text{ and}\ldots x_n\right\} = (1-e^{-\lambda 1})\ (1-e^{-\alpha\lambda 2})\ (1-e^{-\alpha\lambda 3})\ldots(1-e^{-\alpha\lambda n}). \quad (3)$$

Some experimental data are compared with this model in Figure 2. J.L. Baker, of IBM, has built a Markov model for correct detection events directly into a continuous speech recognition algorithm.

In addition to their ability to normalize the results of different experiments, these models to some extent permit one to separate and compare the statistics of language and the statistics of the recognition algorithm. The same models apply to isolated word and continuous speech recognition, except that an extra error contribution from the isolated word boundary detector is needed. There is no apparent reason why continuous speech recognition systems with performance as good or better than isolated word systems in comparable tasks should not be possible. Developmental results approaching the best isolated word techniques have already been reported.



Bottom curve: system tuned to individual talker.

Top curve: talker-independent performance for telephone speech.

Figure 1. Current State-of-the-Art Performance in Isolated Word Recognition

| Event: | Joint false alarm rate predicted by equation (3) with $\alpha$=2.4: | Observed joint false alarm rate: |
|---|---|---|
| $\{A_1, A_2\}$ | 24.7 | 27 |
| $\{A_3, A_4\}$ | 80.8 | 67 |
| $\{A_1, A_2, A_3\}$ | 3.5 | 4 |
| $\{A_1, A_2, A_3, A_4\}$ | 1.5 | 4 |

Figure 2. Comparisons of observed false alarm rates with predictions made by equation (3) for multiple pattern phrases in continuous telephone speech. Observed unconditional probabilities $Pr\{A_i\}$ of false detection in an interval ≈ 0.17 second are: $Pr\{A_1\} = 0.051$, $Pr\{A_2\} = 0.110$, $Pr\{A_3\} = 0.084$, and $Pr\{A_4\} = 0.205$

## ROC Curves

Normalization of different test results frequently requires an estimate of the relation between reject (no decision) rate and correct recognition rate. Relatively few reports give this function, so there is little published information to use as a check on the model. Hence, the following model is not known to apply to anything but the Dialog system. The algorithm produces a goodness of fit score for each decision of interest. Over many trials the fit to a particular reference template has a distribution which is not Gaussian; but the difference between the scores for the closest fitting template and the correct choice template does seem to be approximately Gaussian. A reject criterian based on this function is illustrated in Figure 3. The model yields a family of parallel lines on a probability scale graph; thus only one measurement is required to determine which line corresponds to the system under test. A similar model can be derived for the more commonly used reject criterion in which the input is rejected if no template matches it sufficiently well.

## Probable Error of Measurements

The vast majority of published reports in the field do not contain enough information to establish error estimates for the claimed numerical test results. From this symptom and many others which vary from paper to paper, one may justifiably conclude that the average experimental quality of current speech R&D investigations is absolutely terrible.

To obtain statistical confidence intervals for a parameter, it is necessary to known something about its probability distribution. Most reports contain no helpful information whatever, so one can only guess. For small sized test samples a non-parametric approach can be taken: the experiment is modeled as a series of Bernoulli trials with a binomial distribution of test scores. This method produces seemingly pessimistic estimates of the probable range of random sample test results; but caution and pessimism are the correct attitudes to adopt when interpreting small-sample statistics. Tables of confidence intervals for the binomial distribution are available in an RADC report.

For medium sample sizes (more than 30 trials per talker and more than 30 talkers) a better procedure is to assume that the total number of errors has a Poisson distribution. There is some evidence that the Poisson law is actually a good model for pattern recognition methods which show a low error rate; but the main advantage is that the Poisson distribution has only one parameter, so the answer can be looked up immediately in Figure 4. To use the table, count up the total number n of errors observed in the experiment and find the upper and lower bounds of the desired confidence interval from the appropriate columns. The total number of trials in the experiment is immaterial; the tabulated figures represent total numbers of errors, and must be divided by the

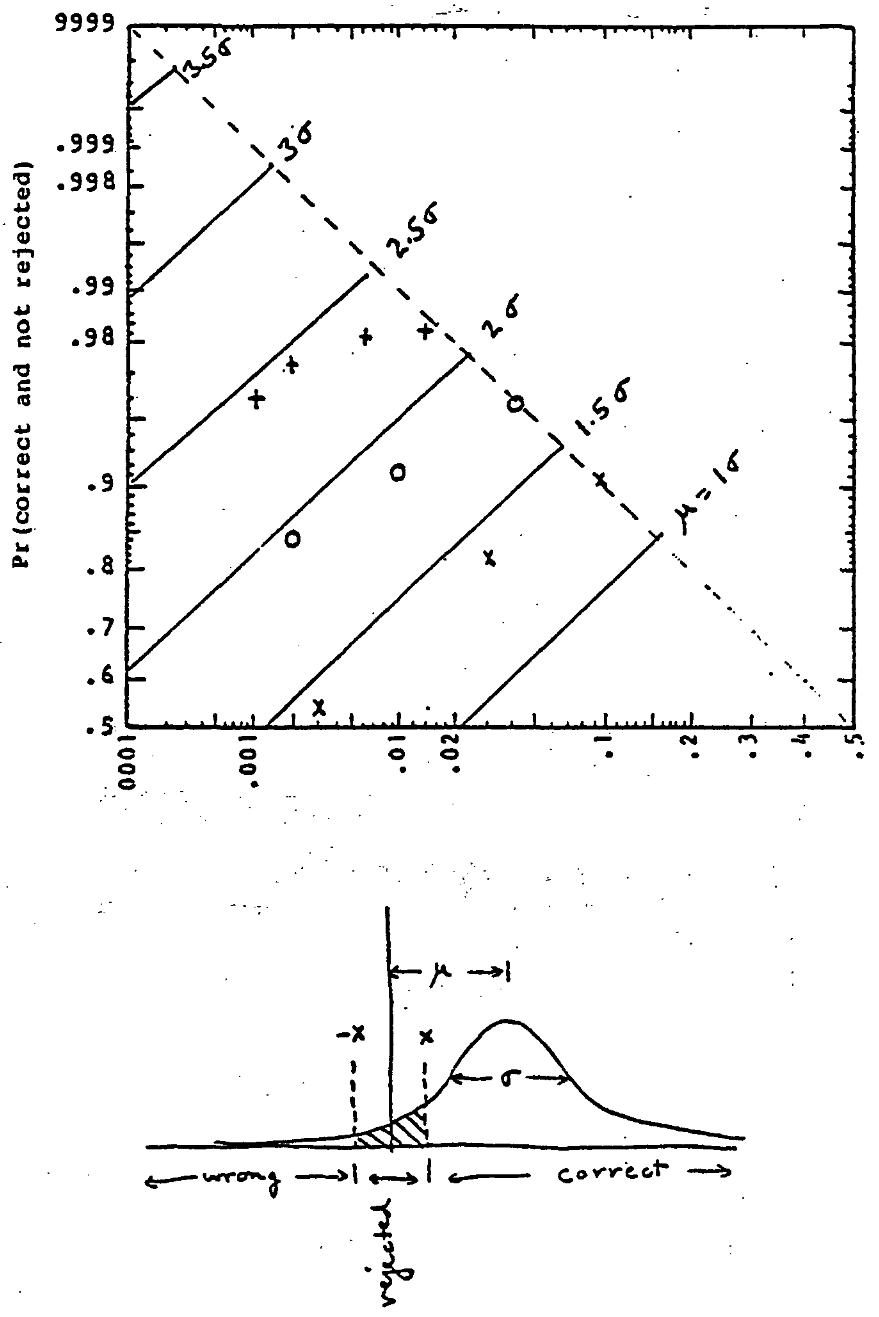


Figure 3. ROC curves for a quasi forced-choice decision rule. Experimental data are plotted for a selected 8-word vocabulary (+), the 10 digits (O), and a 34-word vocabulary (X) in a discrete word recognition task with telephone speech and many talkers.

| n | c = .01 | .05 | .10 | .90 | .95 | .99 |
|---|---|---|---|---|---|---|
| 0 | 0.010 | 0.051 | 0.105 | 2.303 | 2.996 | 4.605 |
| 1 | 0.149 | 0.355 | 0.532 | 3.890 | 4.744 | 6.638 |
| 2 | 0.436 | 0.818 | 1.102 | 5.322 | 6.296 | 8.406 |
| 3 | 0.823 | 1.366 | 1.745 | 6.681 | 7.754 | 10.045 |
| 4 | 1.279 | 1.970 | 2.432 | 7.994 | 9.154 | 11.605 |
| 5 | 1.785 | 2.613 | 3.152 | 9.274 | 10.513 | 13.108 |
| 6 | 2.330 | 3.285 | 3.895 | 10.532 | 11.843 | 14.571 |
| 7 | 2.906 | 3.981 | 4.656 | 11.771 | 13.148 | 16.000 |
| 8 | 3.508 | 4.695 | 5.432 | 12.995 | 14.435 | 17.403 |
| 9 | 4.130 | 5.425 | 6.221 | 14.206 | 15.705 | 18.783 |
| 10 | 4.771 | 6.169 | 7.021 | 15.407 | 16.962 | 20.145 |
| 11 | 5.428 | 6.924 | 7.829 | 16.598 | 18.207 | 21.490 |
| 12 | 6.099 | 7.689 | 8.646 | 17.782 | 19.443 | 22.821 |
| 13 | 6.782 | 8.464 | 9.470 | 18.958 | 20.669 | 24.139 |
| 14 | 7.477 | 9.246 | 10.300 | 20.128 | 21.886 | 25.446 |
| 15 | 8.181 | 10.036 | 11.136 | 21.293 | 23.097 | 26.743 |
| 16 | 8.895 | 10.832 | 11.976 | 22.452 | 24.301 | 28.030 |
| 17 | 9.616 | 11.634 | 12.822 | 23.606 | 25.499 | 29.310 |
| 18 | 10.346 | 12.442 | 13.672 | 24.756 | 26.692 | 30.581 |
| 19 | 11.082 | 13.255 | 14.525 | 25.902 | 27.879 | 31.845 |
| 20 | 11.825 | 14.072 | 15.383 | 27.045 | 29.062 | 33.103 |
| 21 | 12.574 | 14.894 | 16.243 | 28.184 | 30.241 | 34.355 |
| 22 | 13.329 | 15.719 | 17.108 | 29.320 | 31.415 | 35.601 |
| 23 | 14.089 | 16.549 | 17.974 | 30.453 | 32.585 | 36.841 |
| 24 | 14.853 | 17.382 | 18.845 | 31.584 | 33.752 | 38.077 |
| 25 | 15.623 | 18.219 | 19.717 | 32.711 | 34.916 | 39.308 |
| 26 | 16.397 | 19.058 | 20.592 | 33.836 | 36.076 | 40.534 |
| 27 | 17.175 | 19.900 | 21.469 | 34.959 | 37.234 | 41.757 |
| 28 | 17.957 | 20.746 | 22.348 | 36.080 | 38.389 | 42.975 |
| 29 | 18.742 | 21.594 | 23.229 | 37.198 | 39.541 | 44.189 |
| 30 | 19.532 | 22.444 | 24.113 | 38.315 | 40.691 | 45.401 |
| 31 | 20.324 | 23.297 | 24.998 | 39.430 | 41.838 | 46.608 |
| 32 | 21.120 | 24.153 | 25.885 | 40.543 | 42.983 | 47.813 |
| 33 | 21.919 | 25.010 | 26.774 | 41.654 | 44.125 | 49.014 |
| 34 | 22.721 | 25.870 | 27.664 | 42.764 | 45.265 | 50.212 |
| 35 | 23.526 | 26.731 | 28.556 | 43.872 | 46.404 | 51.408 |
| 36 | 24.333 | 27.595 | 29.450 | 44.978 | 47.541 | 52.601 |
| 37 | 25.143 | 28.460 | 30.345 | 46.083 | 48.676 | 53.791 |
| 38 | 25.955 | 29.327 | 31.241 | 47.187 | 49.808 | 54.979 |
| 39 | 26.770 | 30.196 | 32.139 | 48.289 | 50.940 | 56.165 |
| 40 | 27.587 | 31.066 | 33.038 | 49.390 | 52.070 | 57.348 |
| 41 | 28.407 | 31.938 | 33.938 | 50.490 | 53.198 | 58.528 |
| 42 | 29.228 | 32.812 | 34.840 | 51.588 | 54.324 | 59.707 |
| 43 | 30.052 | 33.687 | 35.742 | 52.686 | 55.449 | 60.883 |
| 44 | 30.877 | 34.563 | 36.646 | 53.783 | 56.573 | 62.058 |
| 45 | 31.704 | 35.441 | 37.550 | 54.878 | 57.695 | 63.231 |
| 46 | 32.534 | 36.320 | 38.456 | 55.972 | 58.816 | 64.401 |
| 47 | 33.365 | 37.200 | 39.363 | 57.065 | 59.935 | 65.571 |
| 48 | 34.198 | 38.082 | 40.270 | 58.158 | 61.054 | 66.738 |
| 49 | 35.032 | 38.965 | 41.179 | 59.249 | 62.171 | 67.903 |
| 50 | 35.869 | 39.849 | 42.089 | 60.339 | 63.287 | 69.067 |

Figure 4. Confidence limits on the mean of a Poisson distribution, given a single sample value, n, of the random variable.

number of trials to get the results as percentages. For example, suppose an experiment with good statistical representation results in 100% accuracy. From the table at 0 errors, the average number of errors per experiment over many repetitions of the experiment should be somewhere between 0.010 and 4.605, with 98% confidence.

## Training Sets and Test Sets

A well established empirical fact is that if a pattern recognition machine is tested on the same data base used in training the machine, the results are always better than if an unknown population is employed for the test. The contributions to this bias can be rather subtle, so the safe test procedure involves procurement of a completely new test data base from talkers not previously used in any part of the engineering development project. (On the other hand, development work directed toward a specific application is best done from recordings of real or simulated operational conditions in order to minimize a different kind of bias.)

Surprisingly, there is very little information on this subject in the mathematics literature. Dialog has, therefore, established a modest analytical project to derive expressions for the statistical bias in cases of interest for pattern recognition. One interesting result for maximum likelihood recognition of Gaussian patterns is that the expected test score for an unknown population is pessimistically low when the training set is of finite size. Figures 5 and 6 show the relation between expected training set and test set scores for one dimensional Gaussian patterns. This particular function is of little or no practical value, since all cases of interest are multi-dimensional. The diagrams illustrate, however, that even in this simple case the bias is a complicated function of the population size and the true error rate.

By the law of large numbers, the bias decreases inversely with sample size for a properly designed method. The system behavior as a function of sample size can, therefore, be estimated roughly by taking measurements at two sizes. At every stage of development and field testing, however, the inescapable conclusion is that small scale pattern recognition tests yield very unreliable estimates of large scale performance.

## II. Accomplishments and Planned Development at Dialog

Our company, Dialog Systems, Inc., was formed in 1971 for the purpose of developing and commercializing speech recognition equipment. The concept derived from earlier work engaged in at Listening, Incorporated on marine bioacoustics, acoustic signal processing, and psychoacoustics. The original idea passed through well-known stages of theory, experiment, development, lack of financing, financing, sales and is now

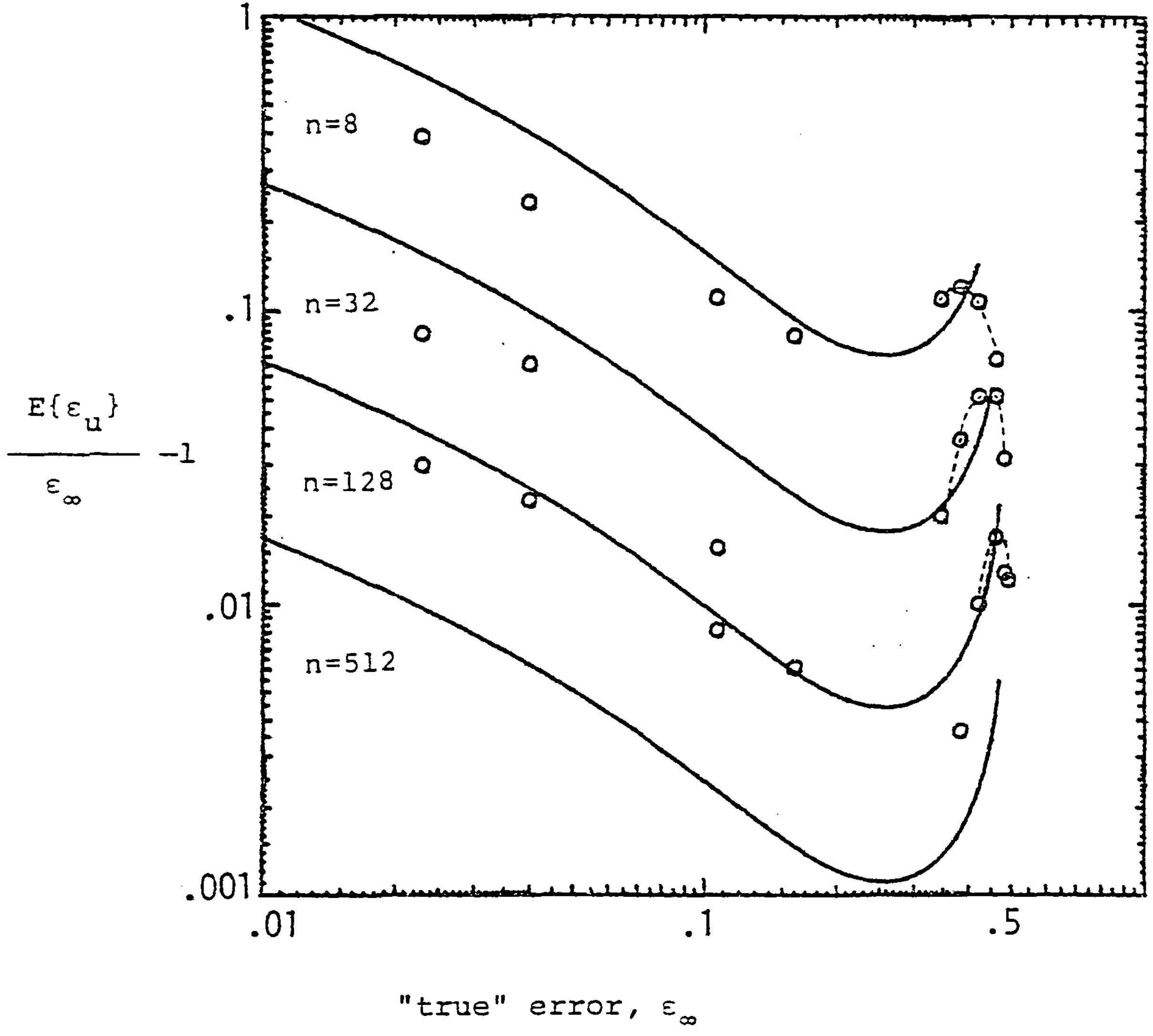


Figure 5. Empirical values for the bias of unknown test set error rates, derived by a computer experiment on pseudo random numbers, compared with a theoretical approximation. The empirical data are relatively accurate (dotted lines) near $\varepsilon_\infty = 0.5$, and the divergence of the Taylor series approximation is evident here. Elsewhere agreement seems rough, but is within experimental error.

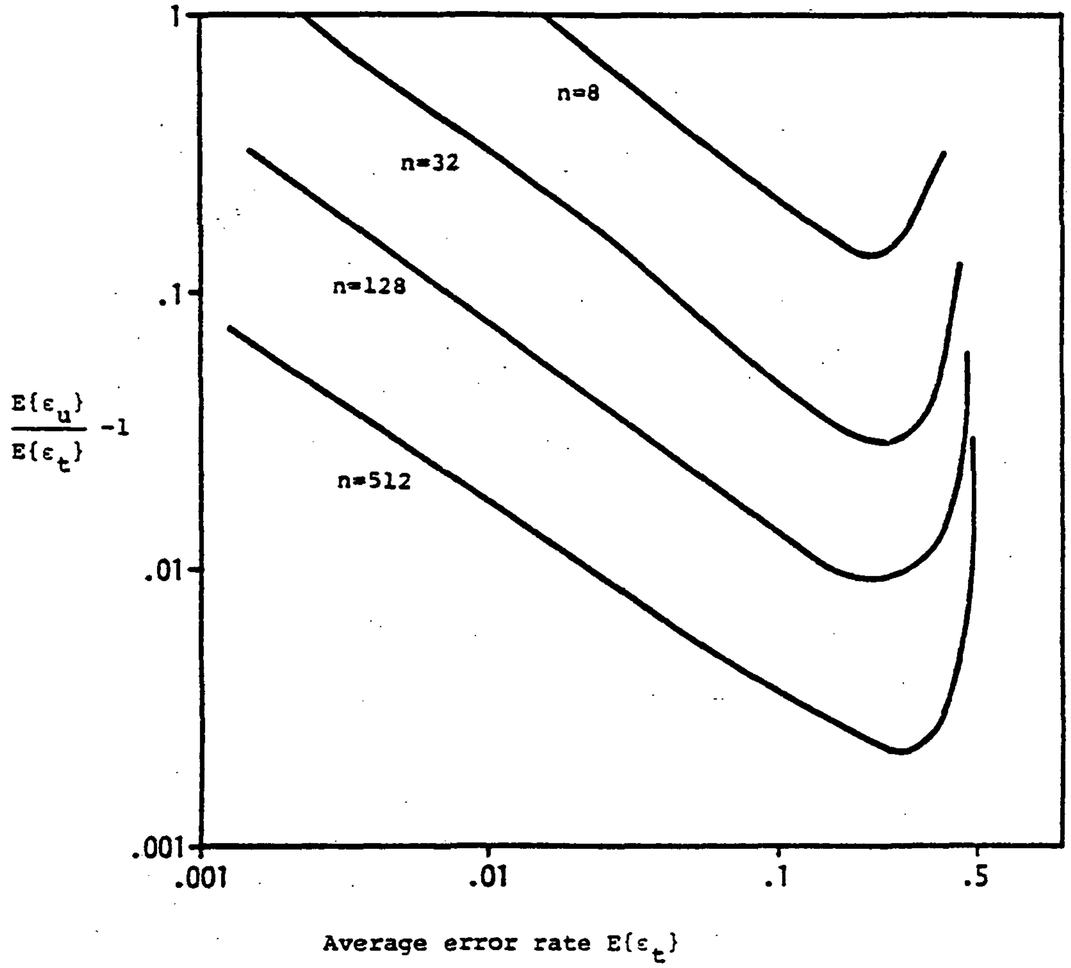


Figure 6. Empirical curves for the relation between average training set and test set error rates for various values, n, of the size of the training set. The functions terminate as indicated near $\varepsilon_t = 0.5$.

The subscript "t" indicates the training set and "u" the unknown test set. The value, $\varepsilon$, is the observed error rate for one experiment of sample size, n.

at the highly advanced state "production engineering headaches". Dialog employs 45, of whom 14 are degreed technical people. The company recently moved from Cambridge to a 20,000 square foot two-building campus complex in Belmont, Massachusetts.

The major product is an eight-channel isolated word system intended for talker-independent switched telephone speech input. With trivial software modification, the same equipment adapts to and tracks each talker's voice characteristics, thus becoming a partially or fully trained machine which is unusually forgiving with respect to changes in the talker's manner of speaking. Operation in the talker-dependent mode requires only one training sample of each vocabulary word. This is made possible by virtue of the precomputed statistical reference patterns contained in the machine.

A complete system (Figure 7) comprises:

1. An analog section consisting of a telephone line switch matrix concentrator, analog-to-digital domain conversion unit and a voice response unit.

2. A disk storage unit for logging and program loading.

3. An interface control computer.

4. A fast signal processing computer of Dialog design and interface to controlled equipment.

5. Power supplies.

The most complex of the units sold to date were priced at about $75,000 for eight simultaneous channels. Installed systems are being supported very heavily by us to ensure that we hear about and correct any troubles encountered. Phone calls from end users are tape recorded and the unintelligible ones analyzed to develop improvements in the recognition algorithm or in the human factors. We have found this operational not-test-but-real-life condition to be different from any simulation, and in the case of new applications to require a substantial refinement effort after delivery and installation. In our experience, problems have arisen that could not be solved by either recognition software or control software changes alone. In general, therefore, the manufacturer must plan for this extra effort, or else the customer must have a speech recognition expert on hand to make his system work. The author has never heard of a speech system working well in an application for which it was not designed, and believes that this situation will continue to be true for some years to come, until a really broad range of application problems has

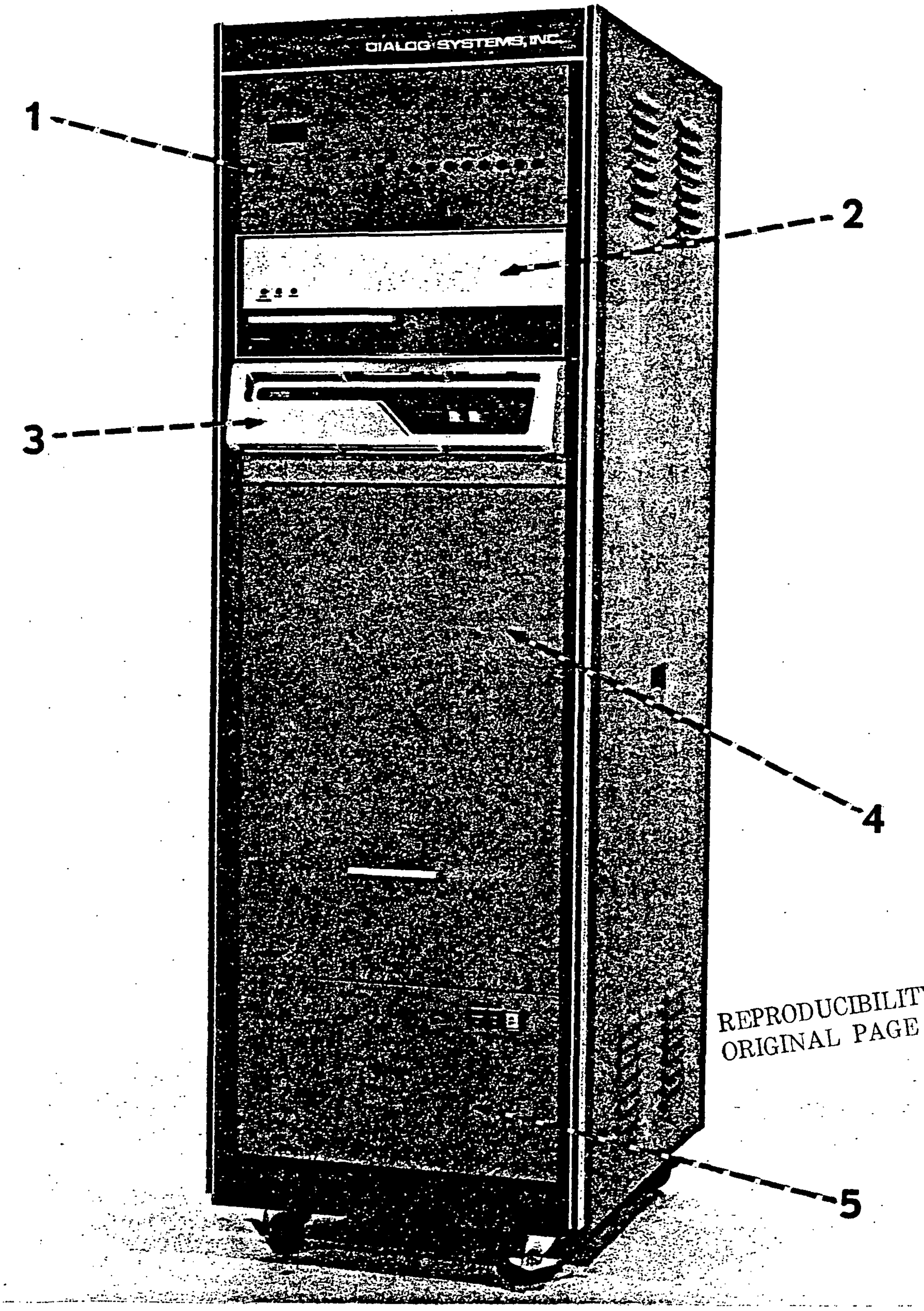


Figure 7. Complete Speech Recognition System

been solved. The problems are being solved, and no one need hesitate to take the next step; but the step is there, cannot be skipped, and it costs money and manpower to take it.

In addition to its heavy investment in product commercialization, Dialog has the resources to maintain a strong research effort; and the company is actively searching for talended people who will follow their own interests in the general area of continuous speech recognition. About 80% of the company's R&D effort is in-house funded, and research personnel (except for the author) are relatively well sheltered from the vagaries of business problems. The main distinguishing feature of the development work at Dialog is that we are making a serious attempt to find improved statistical models of speech data. This includes taking into account the measured variances and cross-correlations of various parameters over large populations of talkers. Thus, we speak of our pattern matching functions as "conditional probability densities" and not "distances". There is, in fact, a fundamental mathematical difference, because a distance is a symmetric relation. Probability measures do not have this property, and do not want it.

Aside from its intrinsically more precise description, an advantage of this approach is that a small number of reference templates suffice for a talker-independent representation. This greatly reduces the workload on higher-level calculations, particularly in talker-independent continuous speech recognition. Our current talker-independent telephone speech product incorporates just two reference patterns per vocabulary word, derived from the speech of hundreds of talkers. The task of gathering, labeling, and proofreading the raw speech data bases for this work has turned into a major project in itself.

While Dialog's engineering activity has so far been devoted to development of system hardware and isolated word recognition, our research effort since 1974 has concentrated on continuous speech recognition. Under contracts with RADC, we have worked on the keyword spotting problem and have produced an algorithm with good talker-independent performance (Figure 8). Word spotting tests under simulated operational conditions are scheduled for 1978. The keyword task is quite difficult, because the brief target sound must be detected independent of context, and all other sounds of an open, plain language input stream must be rejected. The problem is made interesting, however, by the fact that the total number of variables is manageable, so that it is possible to develop theoretical hypotheses and test them by experiment.

The task of limited vocabulary continuous speech has fewer uncontrolled variables than the keyword task, and is therefore easier. Dialog demonstrated such a system in 1975; this system is now in the product development phase and will be released for operational use in 1978.

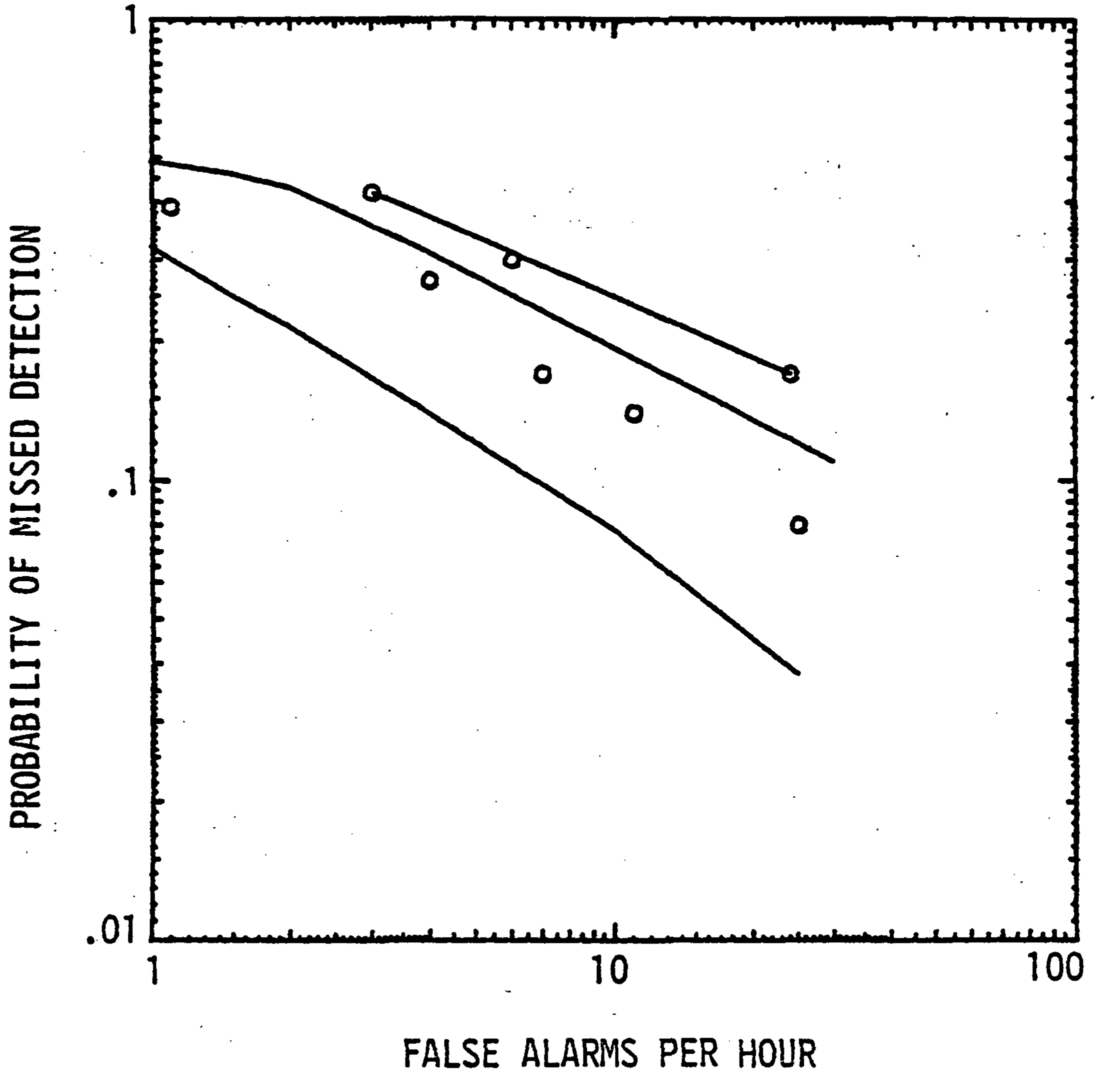


Figure 8. Receiver operating characteristics derived from English language tests of the key word recognition system. Top curve recalls data from the 1976 test, while the two lower curves form an approximation 90% confidence band for the results of the 1977 test.

Organizations which now have operational speech recognition equipment for sale tend to be strong on acoustical pattern analysis and weak on higher level linguistic analysis. There has probably been a general feeling among these people that linguistic processing will not cure the acoustic-level problems on which they feel they are making progress anyway. However, at least two groups, the ones at Dialog and Texas Instruments, have made use of check digits to do error correction on digit string inputs. A series of digits with a check sum is a language; the rule for checking the check digit is a linguistic rule for deciding if a sentence belongs to the language. Thus we already have in practical equipment a rudimentary sort of linguistic processing - and it is not to be scoffed at, because it does reduce the error rate (see Figure 9).

Syntax branching rules have, of course been in use for a long time; but there is still a gap between the well-understood techniques in acoustic analysis and statistical pattern recognition and the realm of linguistic analysis. This gap is being filled in by relatively slow, careful experimental work. It may be expected that practical commercial continuous speech systems with vocabularies of several hundred words will appear in the early 1980's, but probably not within the next two years.

## BIOGRAPHICAL SKETCH

### Stephen L. Moshier

Stephen L. Moshier is President of Dialog Systems, Incorporated, Belmont, Massachusetts, and is specifically responsible for the direction of its research effort in addition to his general administrative duties.

Mr. Moshier has been with Dialog Systems since 1971 and has made major contributions to that company's practical implementation of computerized speech recognition.

From 1965 to 1971, he served variously as Engineering Vice President, Technical Director and President of Listening, Incorporated, Arlington, Massachusetts, where he worked on the development of special purpose transducers and instruments for speech analysis, animal training, underwater acoustics and spectrum analysis.

Mr. Moshier attended Harvard College (Physics), received a Bachelor of Science Degree (Methematics), Summa Cum Laude, from Boston University in 1971 and did graduate work in communications biophysics at M.I.T. in 1971-1973. He has published many papers and patents in the field of speech recognition.

Mr. Moshier is married, has one child and resides in Cambridge, Massachusetts.

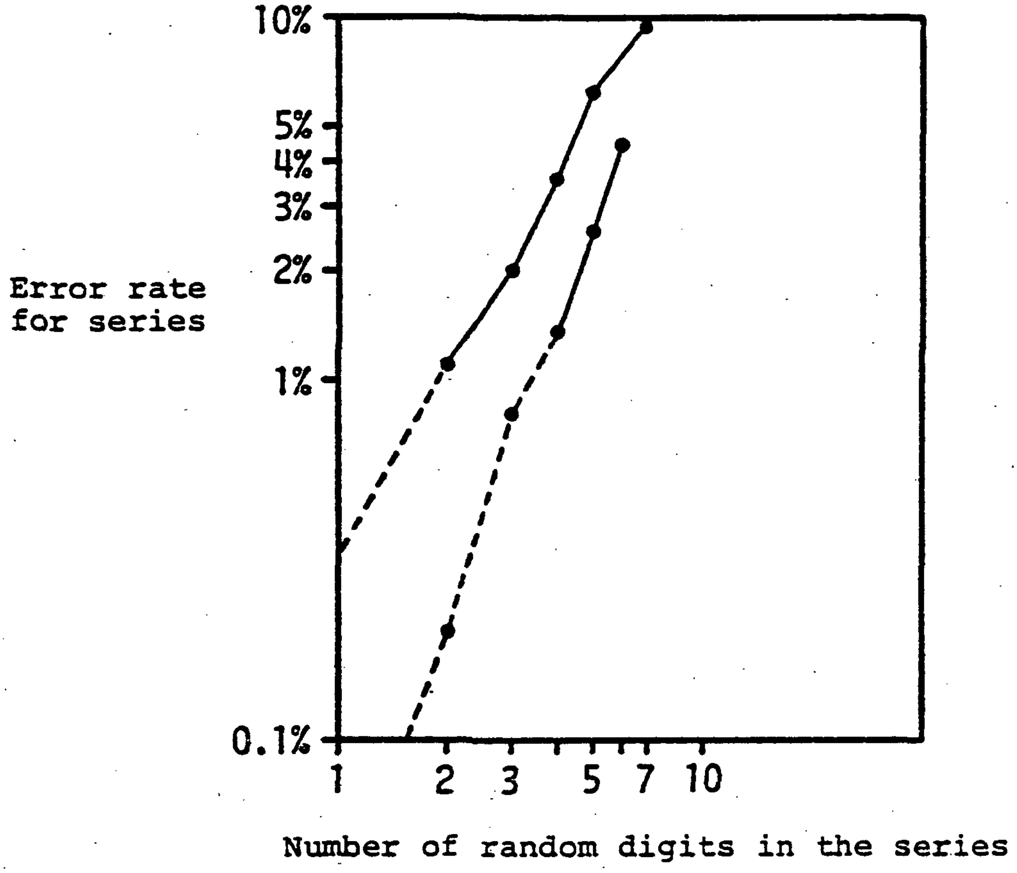


Figure 9. Observed average recognition error rates for random series of isolated spoken digits having check digit(s) appended. For each input series the computer chose the series possessing the best cumulative likelihood score and also having legal check digits. The curves show speaker-independent performance for 25 male voices recorded over standard switched public telephone connections. Error rates below 1% were statistically unreliable in this experiment. No errors were observed for series of length 1.

Upper curve: one check digit adjoined to the indicated number of random digits. The check digit is the 9's complement of the modulus 10 sum of the random digits.

Lower curve: two check digits adjoined. The first check digit is the same as above. The second check digit is the 9's complement of the modulus 10 sum of the squares of the random digits.

## DISCUSSION

## MR. Robert Osborn

Q: Arnold Popky, Threshold Technology: You talked about voice recognition over the phone lines. I was wondering if you felt any of the constraints of the bandwidth of the phone lines and the microphones supplied by the telephone company on your recognition accuracy.

A: That is an area that we have good confidence in at this point. Our algorithm is made purposely transparent to that a limited bandwidth which was never a problem, we never did rely on wide bandwidth, high fidelity speech. And the dynamics of the microphone are adequately taken care of by the statistical approach used. The actual data base is collected. It's a tremendous job to collect adequate data base. We have a data base department that does nothing but record voices, label them, and digitize them. It consists of five full time people. They just collect voices. And it's not easy to collect voices over the random telephone lines. It takes a good deal of effort. We now have data bases consisting of many hundreds of people -- individuals speaking over random telephone lines. It took quite a while to get that. The base recognition accuracy that you saw on the slide for the dialogue speaker independent system, was telephone speech. It was over telephone lines.

Q: Michael Nye: You explained two applications, one was a radio paging application, and the other application for telephone switching. I have two questions. One is, it wasn't clear to me, the benefit that speech offered in that application. And I was wondering if you could just comment for 15 seconds about what that is, why is speech used in that application. And secondly, you outlined a standard system configuration and showed a picture of a device. What would an 8 terminal system like that typically cost, if you can quote a number like that.

A: The answer to the second question is somewhere in the range of $80K. The first question, well, what other types of solutions are there? Yes , you can have a bank of operators listening to people, you can have people touchtone things, or you can have N telephone numbers for N number of people with pocket pagers. Each of those has several economic or operational problems. In the case of touchtones, they don't exist extensively. Installed touchtone base in New York is maybe 15 or 20 percent for instance. In some places they don't have touchtone at all, effectively, Touchtone pads don't seem to work adequately, and they're an additional capital expenditure, much more so than the system that we've presented. Operators are very expensive, especially in places where labor costs go out of sight. Our commercial is in Canada and they're locked-in because they can't get telephone numbers from the telephone company, and if they could, the telephone company would charge them $30 a month for them. And that five or ten thousand subscribers adds up. This is a true economic application area; somebody wrote down on the balance sheet what the results would be, and they came out with voice, I think that's the way we have to approach a lot of these application areas, we've just got to solve the problem.

Q: George Doddington, TI: First, the simple question, what were the key words that were used in that plot of key word performance?

A: That's available in the Rome report. There were a number of key words tried, not only in English.

Q: George Doddington, TI: Second is a comment, that is that, you mentioned a little formula of $x^n$ for performance as a function of vocabulary, and I don't really disagree with that. But I think I would like to make the comment that the performance depends more on the vocabulary than it does on the vocabulary size, and as an example, I would say that we at TI have done some work on nested vocabularies, say, from 100 to 800 words, and we've found, for example, that the performance on the 100 word vocabulary, which are most commonly used words, is poorer than the performance on the 800 word vocabulary, which includes the 100 word vocabulary.

A. We saw that with the AMES group, too. Also another fact that seems to be rediscovered constantly is that the errors are very heavily concentrated on some speakers, they are not uniformly distributed over all speakers, I don't think there is an adequate explanation of why tha's true. We're certainly investigating it. It's an observation I believe other people have made at times, too. It's not related to stress, necessarily.

Q: Ed Huff, NASA Ames: How do you account for the fact,I guess, that your 800 words is dealt with more competently that the 100 word subset.

A: George Doddington, TI: The 800 words are dealt with in exactly the same way as the 100 words, it's just that the extra 700 words that you throw in are more easily recognized. There's this shorthand principle I guess, in speech, that the more often you use a word the shorter it tends to get over eons of time of language evolution, so that the most commonly used words in English are one syllable words like "the," "of," "and," so for example, take the first 100 words. It may be an average of one and a half syllables per word. But after you get beyond the first several hundred words the average number of syllables per word is up around two, and, as everyone knows, two, three and four syllable words are very easy to recognize; the problem is with one syllable words.

Q: Ed Huff: So in other words, you're taking advantage of apertures in order to obtain that result.

A: George Doddington, TI: Oh, yes. We count on the fact that in the exercise of the 800 word vocabulary the first 100 words account for only one-eighth of the exercise. Even if you weight the first 100 words according to frequency of usage, given that they're used more, the results are still the same.

Q: Rex Dixon, IBM: I think one of the things that's happening here is really unfortunate, and that is that we're tending to go to generalizations. For example, the generalization of difficulty of recognition as vocabulary gets larger, with no conditionals, which, of course, as you've pointed out, George, this is a misleading statement at best. I think also your statement about as words get longer they get easier to recognize, is also a generalization. I think any of us, who have been in the speech area, can come up with a vocabulary of long words that will be extremely difficult to recognize, that is to get accuracy within that list, and at the same time come up with a set of very short words that are very easy to recognize. So I think the thing we need to do here is to stop this perpetration, or perpetuation of over generalizations which keep the field in trouble all the time. People go around saying, "Well, as the vocabulary gets bigger, it gets harder to recognize"; "longer words are easier to recognize than short words," etc, etc. And it just simply isn't true. I mean, these things are all conditioned by a lot of other variables. Now the thing we should be about, relative to vocabulary and difficulty of recognition, is saying things like, "here is a method by which you can calculate, using what we know about difficulty, having to do with phonetic similarity, with vocabulary size, here is a way of predicting the difficulty of a particular vocabulary, using all these factors." This is the basic research I hope you were referring to. If we had these things, I think the task for application selection would made easier.

A: I think that's correct.

Q: Jared Wolf, BBN: Just to go along with Rex's statement, I'd just like to point out for some people that may not be familiar with it, that there was a thesis in Carnegie Mellon by Gary Goodman last year which by no means is the whole attempt, but it's a very good first start in just the direction that Rex just mentioned. People should be well aware that we're looking for applications. I don't think it takes care of everything, but it's a lot better than nothing.

omit,
TO
P.288

# LOW COST SPEECH RECOGNITION FOR THE PERSONAL COMPUTER MARKET

HORACE ENEA & JOHN REYKJALIN

HEURISTICS, INC.
LOS ALTOS, CALIFORNIA

## INTRODUCTION

Human speech is a natural and desirable method of interfacing human beings to computers. Its advantages are immediately obvious. It uses the most natural and widely used form of human communication and raises the computer's ability to that of the human rather than reducing the human's capability to the mechanical form required by the machine. Speech recognition has been an active area of computer science research for at least twenty years and is an active research area today, but its wide acceptance in industrial and consumer areas has been hindered by the prohibitive price of the necessary computer equipment. The recent introduction of personal computers based on microprocessors, and the rapidly growing hobby computer market have made it possible to acquaint many people with the possibilities and problems of speech recognition at a reasonable cost.

There are over 100,000 computer hobbyists in the world today and about 40,000 have their own computers. The most popular of these employ the S-100 bus and use the 8080 microprocessor developed by Intel. The average computer hobbyist in the United States today has about $1500 invested in computers and peripherals. A typical system consists of an 8080 based microprocessor, 16K bytes of memory, an alpha CRT, and keyboard. Secondary storage is provided by an audio quality cassette tape recorder, although mini-floppies are becoming popular. While such equipment may seem primitive compared to the sophisticated computers available today in a major university, one can gain perspective by comparing this typical 8080 based computer to the popular IBM 1401 computer.

Typical 1401's had 12K characters of memory with an 11 microsecond cycle time. A typical 8080 based personal computer has 16K bytes of 500 ns second memory (although the processor effectively limits its speed to the equivalent of about 2 microseconds). Thus, while the personal computer lacks the reliable and extensive secondary storage of the 1401, as well as the fast line printer, the CPU is fully as capable as commercially used processors of just a few years ago.

## EDUCATION

In addition to the prohibitive price of a large computer facility, amateurs are not doing research in areas such as speech recognition because of the lack of an elementary introduction to speech recognition techniques. While most hobbyists have a technical background, they are unable to understand the highly specialized papers which are likely to appear in speech journals. Therefore, it is also necessary to provide a considerable amount of material to bridge the gap from the hobbyists background knowledge to that necessary to do meaningful research. Heuristics' Speechlab contains 375 pages of documentation designed so that someone with a high school education can learn the basic ideas of speech recognition. After mastering this material the student should be able to perform his own original experiments.

## THE HARDWARE

Figure 1 is a block diagram of the Speechlab hardware manufactured by Heuristics and sold through personal computer dealers for $299.00. The audio input is amplified and passed through three band pass filters that encompass the ranges 150 to 900 Hz, 900 to 2.2 KHz, and 2.2 KHz to 5 KHz. These ranges, of course, roughly correspond to the frequency ranges of the first three resonances of the human vocal tract. The output of each filter is then passed to a time averager. In addition a zero-crossing detector generates a voltage proportional to the number of times the raw waveform crosses the rest level in a given period of time. Both logarithmic and linear amplifiers are available and one can be selected under software control before the signal is passed to a 6-bit A/C converter. It is also possible to bring the raw waveform into the computer bypassing all of the filters. The board contains hardware reference generators to permit on-board calibration and test. This is particularly important since many of the units are built from kits. Speechlab occupies one slot of an S-100 bus, and the output of the A/D is directly fed onto the computer data bus.

Six bits of information are adequate to characterize all of the parameters measured by the preprocessor. With four quantities being measured at 100 times a second, the data rate from the preprocessor to the computer is 2400 bits per second, well within the processing rates of typical personal computers.

## SOFTWARE

While Speechlab is a laboratory and capable of use in many different configurations, Heuristics does supply a program that will recognize up to 64 isolated words after training by an operator.

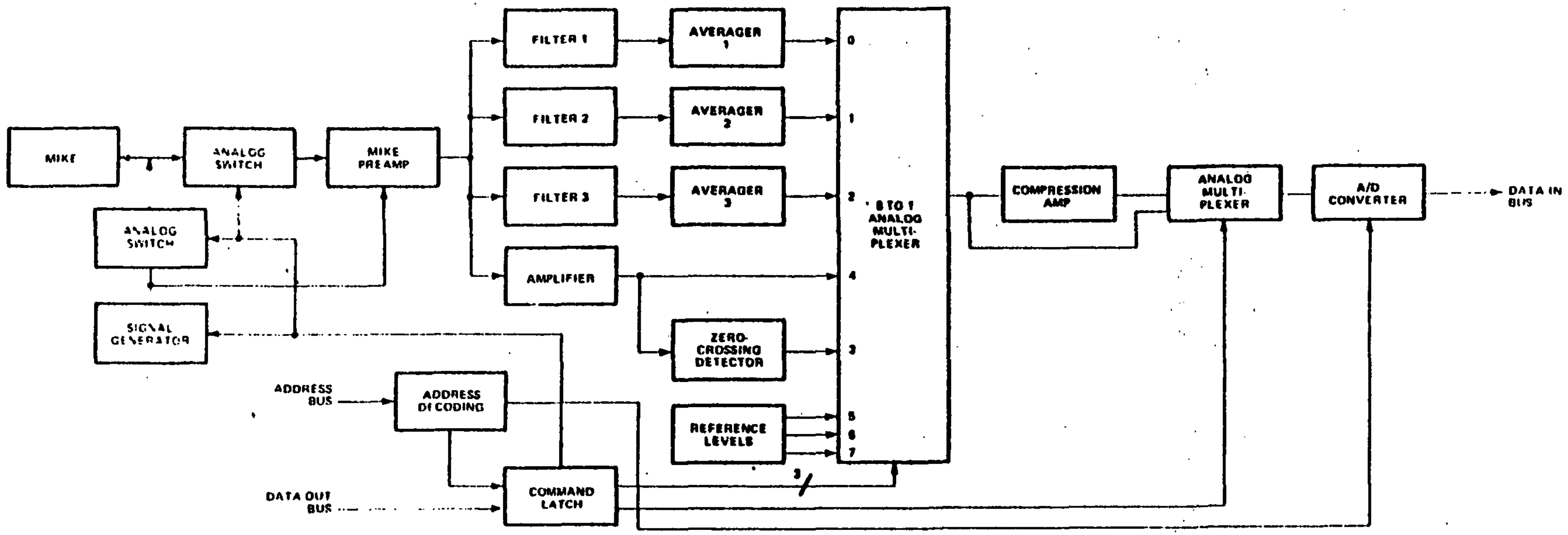


Figure 1.  Block Diagram

The bandpass filter technique greatly simplifies the problem of determining the beginning and end of an utterance. Summing the output from the three bandpass filters results in a measure of the energy in a 10 ms period. Comparing this energy to a threshold computed to be greater than the sum generated by background noise is usually sufficient to accurately determine the beginning of a speech utterance.

The determination of the end of the utterance uses a similar technique. The summed energy from the three bandpass filters falling below a threshold for 100 milliseconds signals the end of the utterance. This criteria allows for stops within a word of up to 100 milliseconds duration.

## NORMALIZATION

Once the beginning and end of the utterance have been found the algorithm normalizes the length by dividing it into 16 equal parts, interpolating between samples if necessary. This time normalization results in 64 bytes of data for each word to be recognized.

Amplitude normalization is accomplished by computing the average amplitude of the utterance and translating the utterance so that its arithmetic average is always the same.

## DISTANCE MEASURE

One training sample is used for each word to be recognized. During recognition the unknown template is compared to previously stored reference templates by subtracting corresponding samples and accumulating the absolute differences (Chebyshev norm). The sum is computed for each word in the table and the entry with the smallest difference is chosen unless the minimum difference exceeds a preset rejection level. While squaring the differences before summing produces better results, the saving in computation time justifies the use of the Chebyshev norm. The algorithm described here, while simple, achieves acceptable recognition in the high 90 percent level while keeping the costs low.

With the advent of the personal computing speech recognition laboratory the number of people actively engaged in speech recognition research will increase by an order of magnitude, and the number of different backgrounds brought to bear on the problem will likewise increase. The home computer hobbyist is as well equipped as the university researcher of just a few years ago, and there is nothing to prevent him from making significant contributions to speech recognition research.

D15

[redacted]

N93-72627

# SPEECH SYSTEMS RESEARCH AT TEXAS INSTRUMENTS

DR. GEORGE R. DODDINGTON

TEXAS INSTRUMENTS INCORPORATED
DALLAS, TEXAS

515-32
176351

## I. TI Capabilities

Texas Instruments supports a Speech Systems Research branch in its Central Research Laboratories. The charter of this branch is to foster the development of new TI business opportunities through development and application of automatic speech processing technology. Seven speech research programs are currently active: Two corporate funded programs determine the strategic direction of our speech research; these are programs to develop automatic dictation technology and low-cost vocoder technology. Three programs are externally funded by Rome Air Development Center to develop and apply advanced but near-term speech processing technology. These programs are: "total voice" speaker verification, limited vocabulary continuous word recognition, and automatic language identification. Finally, two programs are supported internally by TI's operating divisions.

The Speech Systems Research Computer Laboratory contains a variety of computer systems for speech research, system evaluation and product development. System 1, the principal research system, is diagrammed in Figure 1. The salient features of this system include real-time speech I/O, a 500 Mbit disk, a Floating Point Systems AP120B array processor, and a Tektronix interactive graphics terminal with hard-copy.

## II. TI Achievements

### A. Voice Authentication

Although Texas Instruments has been active in a variety of speech processing problems including speech analysis, speech synthesis, word recognition and speaker verification, most of our research effort in the past has been devoted to the development of speaker verification technology. Speaker verification, in its operational format, we refer to as voice authentication. A sequence of 3 programs, funded by RADC and beginning in 1972, led to the development of a voice authentication technology capable of meeting

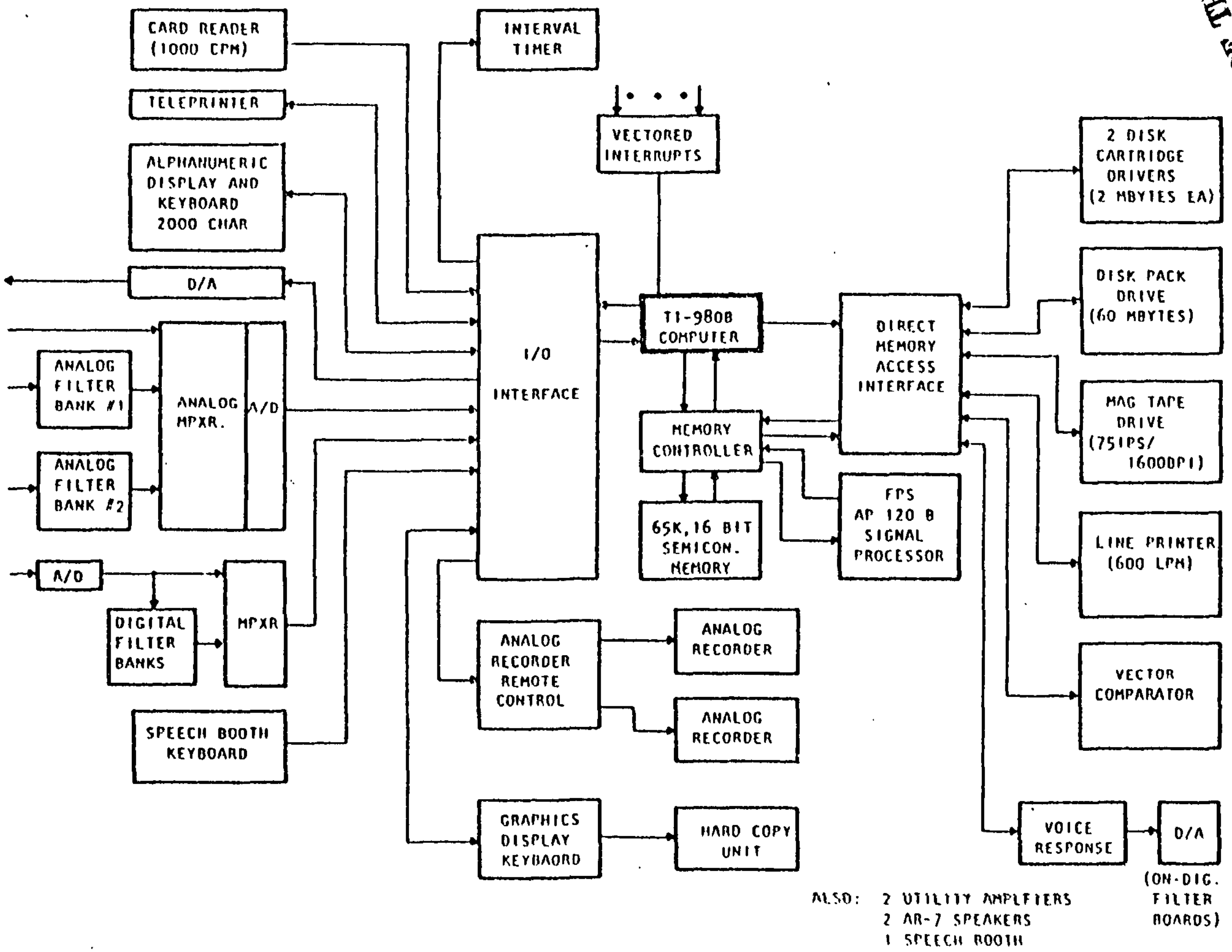


Figure 1. Functional Diagram of the TI Speech Research Computer System #1

BISS[1] performance requirements of less than 1% user rejection at less than 2% impostor acceptance. In retrospect, there were 3 primary problems that we solved in this development:

1) Enrollment

   Enrollment of a user on the voice authentication system must be performed in a single session. Enrollment is difficult for the following reason: Speech data collected in a single session is relatively self-consistent but not representative of an ensemble average over many sessions. Therefore, the initial reference data is biased and the initial estimate of speaker variance is invalid. This problem has been effectively solved by requiring extra speech input data in a special 4-session "post enrollment" strategy and, recently, an additional special 8-session "post-Post enrollment" strategy.

2) "Goats"

   Voice authentication system users are classified as either "sheep" or "goats". The sheep are well behaved and far outnumber the goats. The system performs well on sheep. The speech data of goats exhibit high variance, but the voice authentication system must perform well for everyone. Uniform performance is achieved by a carefully designed decision function which requires more speech data from the goats while, at the same time, not prejudicing the verification decision against them.

3) Discipline

   Voice authentication system users have little interest and little interaction with the authentication system. Authentication utterances often have false starts or are imbedded in extraneous speech data. The verification system must be able to extract the proper input data and discriminate between proper data and garbage. Time registration

[1] Base and Installation Security System, a program administered by the Air Force to define and develop future military security systems.

through energy end-points cannot solve this problem. Proper time registration is achieved through a continuous spectral matching algorithm.

Texas Instruments has controlled access to its Corporate Computer Center by voice authentication for the past three years. This system is in operation 24 hours/day and has provided over ½ million verifications. Current system performance is ½% user rejection at an impostor acceptance level of 1½%. Most user rejections are attributable to noncooperative quirks of the user.

B. Word Recognition

Texas Instruments has operational real-time demonstrations of word recognition for isolated and connected words, enrolled and independent speakers, and small vocabularies up to 50 words. Large vocabulary recognition has been performed in non real time for vocabularies of greater than 1,000 words. Also, the development of automatic language identification has been sponsored by RADC. One significant result in language identification is the demonstration of improved identification by normalization of long-term spectral averages. Although long-term spectral averages have been shown to be useful in discriminating languages, the wide variety of recording conditions encountered in our data base invalidate such utility. With the incorporation of spectral normalization, results of 5-language identification task have been improved to 80% correct on excerpts of two minutes duration.

"Total Voice" speaker verification involves two speech processing tasks: speaker independent recognition of an identifying sequence of six spoken digits; and speaker verification using the user identification and the same identifying speech input data. The application requires speaker independent recognition of a connected sequence of six digits with less than 1% sequence recognition error. These severe application requirements have been achieved by incorporating two check digits in the sequence for improved recognition accuracy and by "forbidding" certain digit pairs such as "three-eight".

## III. Fundamental Problems

I have ordered below four classes of problems that must be faced in the development and deployment of an automatic speech processing system:

A. Speech Science

The lack of speech science limits the capabilities of speech processing systems. So, how do we go about getting speech science? Careful direction is exceedingly important because one can easily drown in the vast, uncharted oceans of speech phenomena. In my opinion, a very good way to get speech science is to identify an important real application for speech processing and to persevere in developing speech technology for this application. This approach, which I refer to as the "correct" approach, is contrasted with the traditionally popular method in Figure 2.

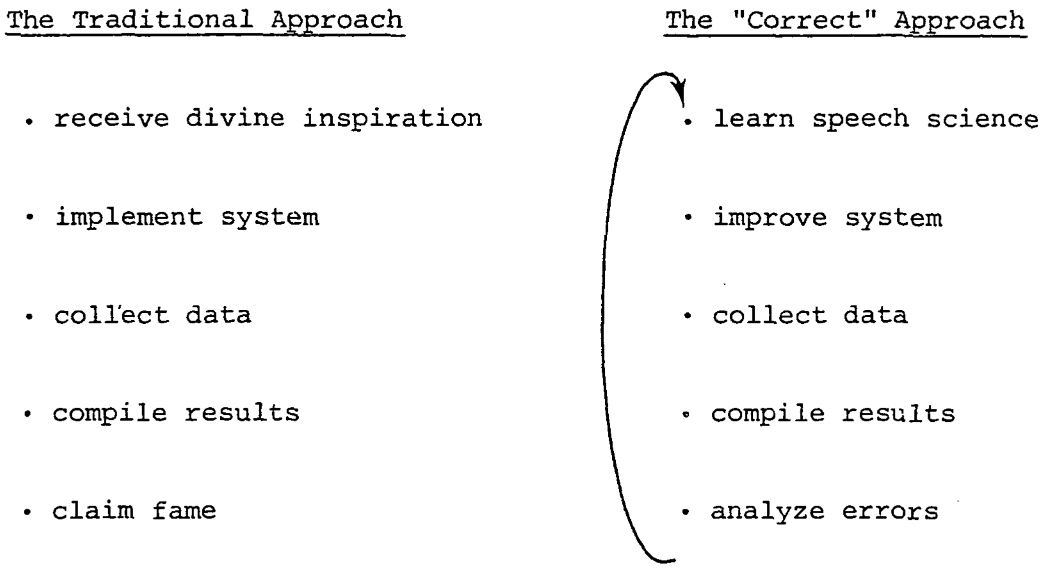
### B. Cost

Cost is a very important consideration in automatic speech processing for two reasons: First, speech processing is a complex problem which is inherently costly, at least in terms of computing power. Second, system cost effectiveness is usually measured in terms of the efficiency of a person, at least in the case of speech recognition. Cost/benefit tradeoffs must be carefully made between speech and other alternate media.

### C. Performance Forecast

It would be nice to do an experiment in the laboratory and be able to say with confidence that the laboratory results will be realized in the operational system. This rarely, if ever, happens because the laboratory data is not representative of operational data. Sometimes the discrepancy between laboratory results and operational results is embarrassingly large. One important reason contributing to this is the typical favorable mix of sheep with goats in laboratory experiments. (Speech researchers are usually "super sheep".) System performance depends very strongly on this mix. A large number of subjects is required to properly evaluate system performance. Figure 3 is included to provide some perspective on the sheep/goat problem. This figure shows a histogram of user data variance for an operational voice authentication system.

How much data must be included in a laboratory experiment to provide a good performance forecast? My rule of thumb answer to this question is that enough data must be collected to provide 30 errors. Assuming that each trial is independent, thirty errors will provide you with an estimate of the true error rate within ±30%, for the given data context, with 90% confidence. Suppose for example that you anticipate 1% error for a certain word recognition system. This implies that at least 3,000 spoken words must be collected to provide the desired confidence interval on error rate. Note however, that the trials must be statistically independent. Will you collect 3,000 words from one speaker or one word from each of 3,000 speakers?



The Traditional Approach

- receive divine inspiration
- implement system
- coll'ect data
- compile results
- claim fame

The "Correct" Approach

- learn speech science
- improve system
- collect data
- compile results
- analyze errors

Figure 2. Basic Steps in Developing Speech Technology

REPRODUCIBILITY OF THE ORIGINAL PAGE IS POOR

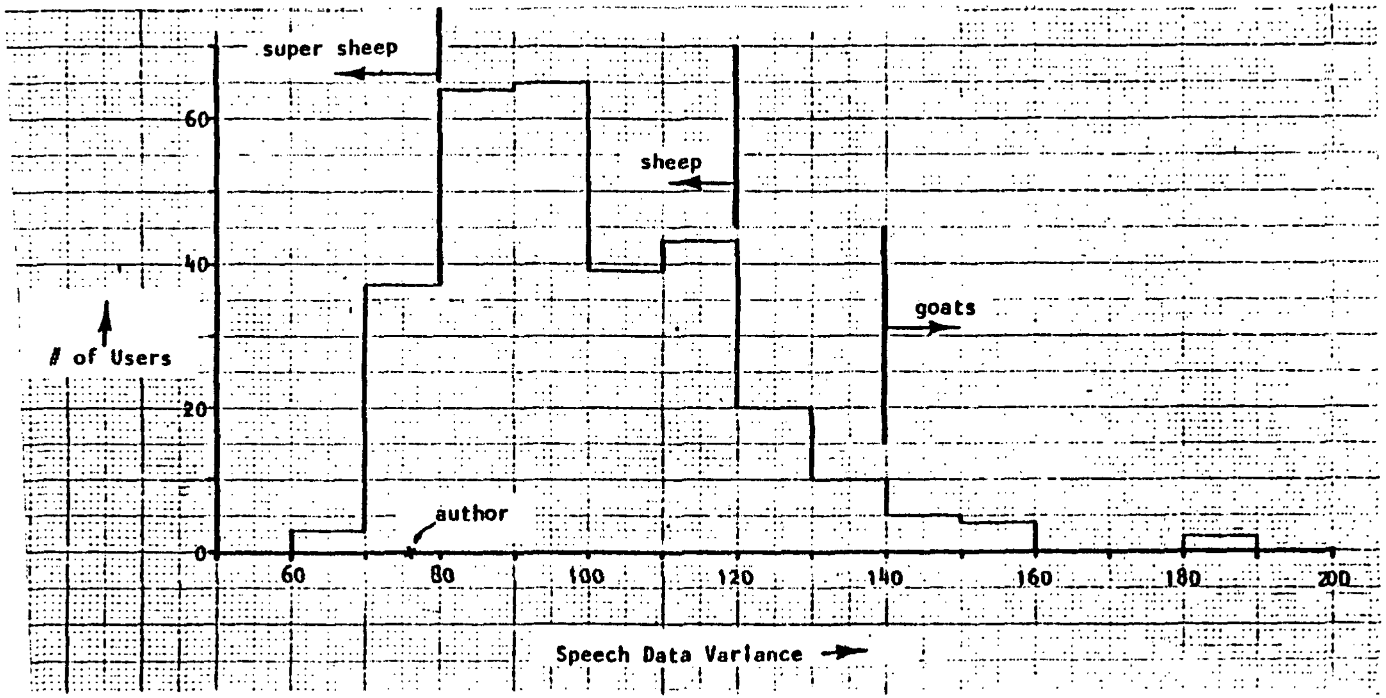


Figure 3. Histogram of User Data Variance for Operational Voice Authentication

### D. The Human Factor

User acceptance is a critical factor in speech processing systems. For word recognition this includes not only recognition performance, but also other recognition characteristics. For current technology, isolated word recognition machines, a most important operating characteristic is the requirement for pausing between words. This is a nontrivial skill to perform reliably and is a major underlying factor in initial performance degradation. Fortunately, system performance is often aided through the adaptation of the user to the system. This includes learning to speak clearly and loudly to the system in spite of the fact that the microphone is often less than 1 inch from the mouth. Loud, clear input stabilizes the speech data and improves system performance. Such user adaptation is clearly demonstrated in Figure 4. Figure 4 is a plot of user data variance as a function of session number for an operational voice authentication system. The subtle feedback in this system has provided a user learning time constant of 2,000 sessions.

## IV. Technology Forecast

Progress in speech systems development is tied closely to developments in computer technology. Advanced speech system capabilities will require inexpensive yet highly competent high speed data processing. The speech processing system will comprise a general purpose CPU with speech input through a special purpose speech preprocessor and feature extractor. An important cost element is this speech preprocessor unit. TI is currently developing, under contract with ARPA, a one-chip speech analyzer using CCD technology. This chip implements a 19-channel sampled data filter bank with on-chip 4-bit A/D conversion and multiplexing.

Low cost speech preprocessor technology coupled with advances in microprocessor performance is anticipated to have substantial impact on speech system competence, cost and market size by 1980-1982. At this time useful and affordable capabilities will be introduced for connected word recognition and narrowband·digital voice communication.

REPRODUCIBILITY OF THE
ORIGINAL PAGE IS POOR

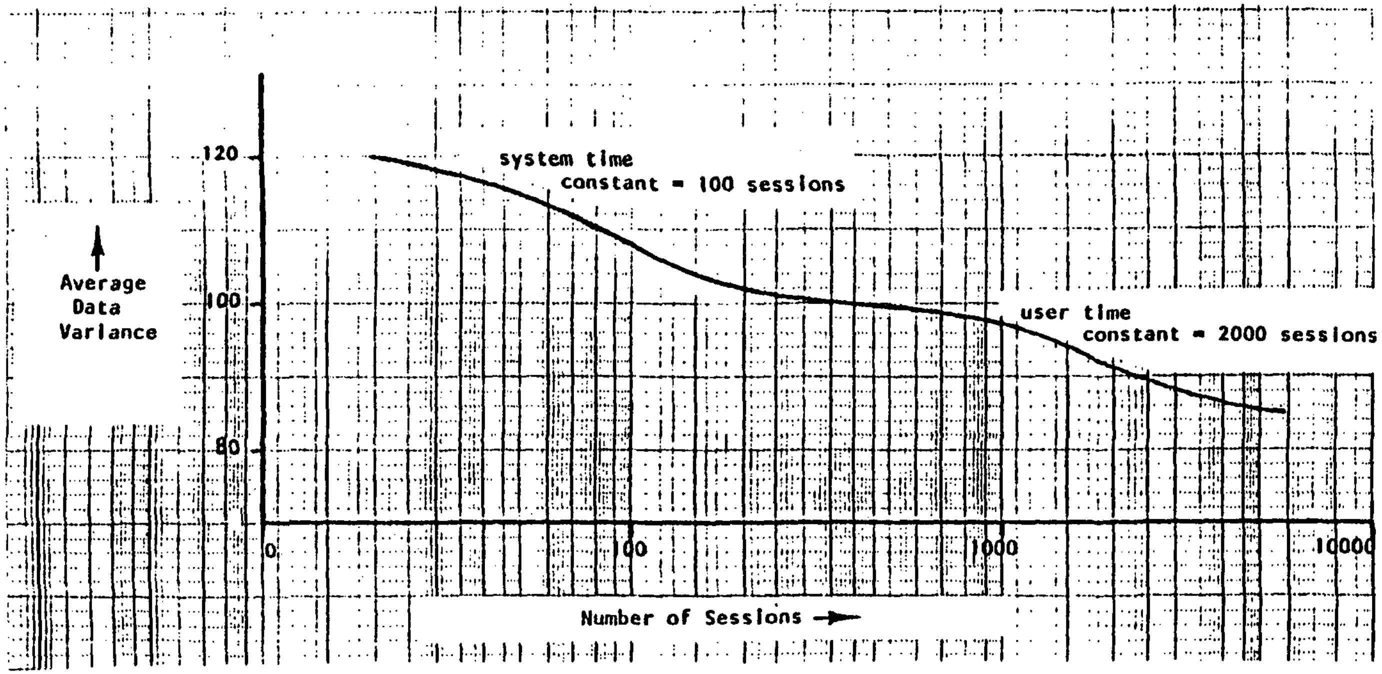


Figure 4. User Data Variance as a Function of Session Number for Operational Voice Authentication

# BIOGRAPHICAL SKETCH

## George R. Doddington

Ph.D. in Electrical Engineering, University of Wisconsin
M.S. in Electrical Engineering, University of Wisconsin
B.S. in Electrical Engineering, University of Florida
Professional Engineer, State of Wisconsin

At Texas Instruments, Dr. Doddington has directed programs of speech research encompassing advanced speech processing techniques. This work has included interactive simulations of word recognition, speech segmentation and analysis, and speaker verification. Dr. Doddington's doctoral study emphasized communication theory, control theory, probability theory, and neurophysiology. His doctoral research was conducted at Bell Telephone Laboratories during 1969 and 1970. This work comprised the development of a method of nonlinear time normalization of speech and the implementation of this method in a system of speaker verification. Dr. Doddington joined Texas Instruments in 1970. Dr. Doddington's master's thesis comprises a generalized theory for relating the gross operating characteristics of chromatographic systems to the statistics of molecular behavior. Dr. Doddington was employed at the Federal Communications Commission from 1960 through 1963, during which time he designed and developed a secrecy coding system for radioteletype communication. In 1964 he received the bachelor's degree with high honors for his thesis comprising the theory and practical implementation of a method of linear amplification approaching 100 percent efficiency.

## DISCUSSION

## Dr. George Doddington

Q: Mark Medress, Sperry Univac: Do you have any feeling for how much action you got out of the check digits in your total voice verification system?

A: Okay, I didn't talk about the performance of that system, really, except I did say we had eight errors in about a thousand trials. Yes, I have some feeling for that and I'll tell you. The eight errors in a thousand trials represented about one percent error. Now that's on the six digit sequences themselves. We have run some experiments using the digit recognizer component, in the sequence recognition strategy, and we've gotten about 95 percent correct digit recognition. Now those two performance figures are not comparable. The digit recognition is for digits, 95 percent correct, and the sequence recognition is for sequences, and that's about 1 percent error.

Q: Mark Medress: That's 95 percent of the words in those sequences are correct, is that what you're saying?

A: That's right. 95 percent of all digits were correct.

(This page intentionally left blank)

D/16

[redacted]

N93-72628

# MULTI-USER REAL TIME WORD RECOGNITION SYSTEM

516-32
176352

S. S. VIGLIONE

SPEECH RECOGNITION GROUP
INTERSTATE ELECTRONICS CORPORATION
ANAHEIM, CALIFORNIA

Interstate Electronics Corporation is presently marketing a discrete word recognition system to be used as a voice data entry terminal and capable of handling one to four users with vocabularies of 250 word per user. The recognizer is an acoustic pattern classifier that produces a digital code as an output in response to the received utterance. It consists of a spectrum analyzer, an analog multiplexer and A/D converter, a programmed digital processor, a reference pattern memory, and an output register.

The spectrum analyzer divides the input audio spectrum into 16 frequency bands that cover the useful frequency range. By means of parallel detection and lowpass filtering the resulting 16 analog signals represent a power spectrum that constitutes the feature for speech classification. These 16 continuous signals are multiplexed, sampled at 200 Hz, and converted to digital form with 8-bit precision. Thus, the original utterance arrives at the digital processor as a string of 8-bit binary numbers.

The coding compressor compensates for changes in the rate of articulation and reduces the spectral data generated by each utterance to a fixed-length code for the classifier. It reduces every word, regardless of length, to a 240-bit pattern. As a result, the fixed-length codes can be processed in real time by simple pattern recognition techniques without the need for a great deal of high speed memory. The compression algorithm is essentially an arithmetic process, preserving all the properties that change during an utterance and eliminating those that remain steady.

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

After the system has been trained to a particular user, each new pattern from the coding compressor is compared with a syntactically determined subset of all the previously learned reference patterns in memory. The classification process matches the patterns bit by bit via a Hamming-distance classifier.

The system, configured with a NOVA 3/12 with 32K words of core, is capable of supporting up to four simultaneous voice input channels in conjunction with a variety of standard minicomputer peripherals.

The heart of Interstate's voice data entry system is the control program that organizes the system so that it meets a particular application need. This specification of the control program is done in a high level language that permits users to write their own application software, or to modify control programs delivered with the system.

Using the VOICE (Voice Oriented in Core Executive) software operating system supplied as an integral component of the voice data entry system, the user can specify such applications specific system parameters as:

> Configuration parameters, including the vocabulary size, number of users, configuration of input and output devices, and the number and size of internal buffers and data arrays.
>
> The dictionary of vocabulary items to be utilized in an application, along with multiplicity of representations for each vocabulary item.
>
> Dictionary of prompt and error messages. These messages can be displayed for the operator as a guide through a complex data entry sequence. Error messages can be used as a key to enable error correction immediately at the data source.
>
> An action structure associating an appropriate system action with each command that is recognized. Actions may range from simply outputting a code associated with a recognized word to executing a complex computer program that is a function of several previously input commands.
>
> A syntax structure that associates subsets of the dictionary with specific functions to be performed in the application. The syntax structure provides a context for the user, and permits the use of large vocabularies without loss of recognition accuracy.

Interstate Electronics Corporation has under development an advanced word recognition system capable of handling up to eight users simultaneously. A common speech pre-processor will multiplex and condition the inputs from each station. The heart of the preprocessor is a single board array processor programmed via firmware to perform Hamming weighting of the speech data and an FFT spectral analysis from 80Hz to 8000Hz. The FFT output is processed to detect the peaks in the first four formant bands every 12.5ms. The energy in the spectral bands is amplitude normalized, and the detected formant energies are processed to provide 16 time normalized samples for each utterance following word onset and word ending detection. In addition to the peaks in the four formant bands, three broad energy measures corresponding to the energies in F1, F2 and F4 are also computed along with the gross energy in the utterance. The sixteen time normalized samples of these eight parameters form the pattern vector for classification. The classification is implemented with a "minimum distance" classifier, where, for computational simplicity, the vector components, rather than the vector itself, is used in computing the distance metric. During training a threshold is established for each pattern which permits the generation of multiple templates per word if required.

This system uses a remote user subsystem with two 20-α/N character displays for operating prompting and message verification. The user subsystem also contains operator controls for train, one-word selectable train, one-pass retrain as well as test and operate functions. The system is configured with a controlling minicomputer and floppy disc operating under DOS for control and application programs.

Laboratory work is underway to extend the discrete word recognition system to handle word strings and phrases; as well as generalized, small vocabulary word recognition.

## DISCUSSION

## S. S. Viglione

Q: Dave Hadden: As I understand the VDET and the Voice software were Scope products and I'm sort of curious as to the relationship in design of the VDET relative to the Army WRS.

A: The early version of the VDETS system that was developed by Scope was continued under Army contract in the implementation of the WRS. There are some changes that have occurred during the last year which are not incorporated in the WRS, particularly in the coding and compression algorithm. The Army version is a disk operating system. The VDETS as a stand alone system is a Link Tape operating system. The modified algorithm encodes the 16 filter outputs into a 120 bit pattern, then takes the one's compliment to form a 240 bit pattern for classification. The speech input data sampling rate has been increased from 100 to 200 SPS, and the training algorithm now uses a variable number of input samples as opposed to a fixed training sample size. None the less the VDETS is essentially the same as the WRS. The changes incorporated in the algorithms this past year have been directed to, and have accomplished, a significant improvement in classification performance.

Q: John Allen: Can you explain why you do the one complementing algorithm? It seems that you double the amount of data and that its redundant.

A: The 'ones' compliment is implemented to aid in the correlation scheme used for word classification. The 120 bit binary pattern, representing the incoming word is inverted to create a second 120 bit pattern which is the 'ones' compliment of the original. The resulting 240 bit pattern is then "ANDED" with the reference patterns for each word. As a result of the ANDing operation, the first 120 bits of each reference pattern will be 'one' bits if and only if each bit was consistently a 'one' bit for all training samples. Bits 121 through 240 will be 'one' bits if and only if each bit was consistently a "zero" bit for all training passes. The total number of 'one' bits form the basis for classification.

D/7

[redacted]

N93-72629

S7-32
176353

# AUTOMATIC SPEECH RECOGNITION TECHNOLOGY DEVELOPMENT AT ITT DEFENSE COMMUNICATIONS DIVISION

DR. GEORGE M. WHITE, PH.D.

ITT DEFENSE COMMUNICATIONS DIVISION
SAN DIEGO, CALIFORNIA

## INTRODUCTION

ITT Defense Communications Division supports the needs of the Government and Department of Defense in voice processing through a wide range of research and development activities. ITTDCD anticipates significant increases in the interest of government agencies for voice processing equipment in the next few years. This increased interest will be promoted through the maturing of signal processing technology. At least an order of magnitude improvement in the performance/cost ratio can be expected within five years due to advances in microelectronics, i.e. the commercial devices that today sell for $10,000 to $20,000 could be manufactured for less than $500 for many vocoding, speech recognition, and speaker verification devices. This lower cost is expected to open up new application areas within the government and defense community.

Considering speech recognition devices, for the task of recognizing a small vocabulary of isolated utterances, spoken by a small number of known cooperative speakers, in a relatively noise-free environment, over a high-quality microphone, the problems are practical rather than theoretical. There are many algorithms that achieve accuracies in excess of 99.0%. However, the relatively high cost of practical devices (typically more than $5,000) is prohibitive for most applications. The cost of such devices could soon be lowered dramatically by technological advances in LSI and charge transfer devices. For this reason, ITT has considerable interest in charge transfer devices and their application to Fourier analysis and bandpass filtering and Itakura LPC analysis. It is felt that recognition systems of the above type could be built for less than $500 per unit (excluding mask-making costs) that would achieve better than 99.0% correct recognition scores for 50 word vocabularies of polysyllabic utterances that differ in more than one syllable. Of course, the actual costs will be strongly dependent on production volume and the above estimate is based on high volume.

For all but the simplest form of recognition mentioned above, the problems are both theoretical and practical. ITT is actively

[redacted]

pursuing research on the practical and theoretical problems in the areas of speech vocoding (bandwidth compression), speaker verification and speech recognition.

## I. PRE-FY78 TECHNICAL REVIEW

In FY77, ITTDCD completed development of the ITT processor: a high speed programmable signal processor. The ITT processor has been programmed to function as an LPC vocoder. It executes an LPC-10 analysis for the encoding operation, the characteristics of which are described below.

### LPC-10 Characteristics

| | |
|---|---|
| Predictor Order | 10 |
| Sampling Rate | 8 KHz |
| Bit Rate | 2400 bps |
| Frame | 22.5 msec (54 bits per frame) |
| Analyzer | Semi-Pitch Synchronous |

| | |
|---|---|
| Low Pass Filter: | 4th Order Butterworth |
| Pitch: | AMDF function with Dynamic Programming (DYPTRACK) smoothing (50 Hz to 400 Hz, 60 Values). |
| Voicing: | 2 decisions per frame based on Low Band Energy, zero crossing count and reflection coefficients RD1 and RC2. |
| Preemphasis: | $Z_n - \frac{15}{16} Z_{n-1}$ |
| Matrix Load: | Covariance (Modified ATAL) |
| Matrix Invert: | Modified Cholesky Decomposition |
| Coding of RCs: | Log Area Ratio for RC1 and RC2 and linear for others. |
| (The Synthesizer: | Uses Interpolation and is Pitch Synchronous) |

The unit has two processors and two memories (a data memory and a program memory). The LPC-10 analysis code uses only 1238 words of program memory and 2900 words for data memory. It is very fast: 10 LPC (reflection) coefficients are generated in 2 msec, and pitch tracking and voicing analysis are performed in 4 msec for each 22.5 msec window. The processor itself weighs 50 pounds, consumes 180 watts of power, and uses about 180 TTL chips.

## II-A. POST FY77 CAPABILITIES AND PLANS

ITT Defense Communications Division has great interest and capability in the automatic speech recognition (ASR) area. ITTDCD is actively seeking government study contracts in ASR, and is also investing several hundred thousand dollars in salary a year of internal funds on research and development in this area. ITTDCD plans to produce an ASR demonstration system early in FY78.

ITT Defense Communications Division has engineering offices and personnel at Nutley, N.J. and San Diego, California. Both of these installations have PDP-11 computers and associated peripherals which are utilized for voice processing research and development. Programs developed at one location are transferred to the other so that both facilities have a total system capability for pursuing research activities at any given time. It is planned that activities in the area of automatic speech recognition will be supported by both ITT facilities, with a majority of the work performed at San Diego.

The equipment, facilities, and personnel presently allocated to automatic speech recognition are the following:

In San Diego, California, ITTDCD has a 2,000 square foot office with 8 scientists and engineers (3 Ph.D's, 5 senior engineers and programmers) and is acquiring a PDP-11/60 with 96K core, and 14 and 88 megabyte disks, tape drive, 3 interactive "smart" CRT terminals, a graphics display unit with hard copy, and UNIX operating system. The principle activity of the San Diego office is ASR research. The personnel in San Diego includes: Dr. George White, formerly of Xerox (7 years research in ASR); Dr. James Dunn (10 years experience in "voice processing"); Mr. Robert Wohlford, formerly with NSA (10 years experience in ASR research); Dr. A. Richard Smith (Ph.D. in computer science from Carnegie-Mellon University whose thesis is on speech recognition); Mr. Russell Lemon, formerly with the USAF (10 years experience in modems and digital voice processing); Mr. John Lowry (formerly with RAND), a recent Masters level graduate from Carnegie-Mellon University (at CMU he worked on CMU's speech recognition project); Mr. George Vensko, formerly with Technology Service Incorporated (6 years experience in signal processing/speech compression), and Mr.

Douglas Landauer, formerly with Pattern Analysis and Recognition Corporation (two years experience programming in signal processing areas).

In Nutley, New Jersey, a comparable computer facility exists with a PDP-11/55 plus peripherals. The group in Nutley is headed by Dr. Marvin Sambur, formerly of Bell Labs (Dr. Sambur has more than 5 years experience in ASR research and is one of the better known personalities in the ASR field. Directly supporting Dr. Sambur in the ASR work will be: Mr. Paul Gilmour (5 years experience in voice processing with his Masters thesis concerned with Formant Tracking); Dr. Walter Fan (extensive experience in developing software utilizing ITTDCD's disciplines of Top Down design and test and structured programming); and Mr. Anthony Russo (headed ITTDCD's team which developed a hardware version for the Navy of Itakura's LPC synthesizer).

## II-B. POST FY77 DETAILED RESEARCH PLANS

The following are brief descriptions of several of the areas ITTDCD will be investigating in FY78.

### A. DYNAMIC PROGRAMMING

ITT's dynamic programming research has several different aspects. One aspect concerns constraints on the amount of nonlinear time warping that can be performed by dynamic programming. The goal is to insure that the degree of time warping is commensurate with experimentally observed time axis deformations.

It is known that different pronunciations of the same word result in different segmental durations, and that a nonlinear time alignment strategy must be used to match such utterances against standardized templates. However, the degree of temporal variability that should be permitted is not known. For example, some phonemes are characterized by their time rate of change while others are not; e.g., stop consonants are and most vowels are not. It would seem that a time alignment strategy that reflects this fact ought to be better than one that treats an entire utterance with constant constraints on non-linearity regardless of the types of phonemes found in the utterance. Perhaps piece-wise linear matching would be adequate or perhaps strict linearity would be inadequate even for segments as small as phonemes. Research into this problem has been partitioned into the following tasks:

- SPEECH DATA BASE ANALYSIS - to measure the extent to which subword segments exhibit differing degrees of temporal deformation;

- ALGORITHM DEVELOPMENT AND TESTING - to discover computationally efficient means of encoding and utilizing information about the amount of temporal deformation for different speech segments during the classification process:

- AUTOMATIC TEMPLATE GENERATOR DEVELOPMENT - to create a procedure which automatically generates templates that incorporate segment controlled deformation parameters, as well as the usual spectral information;

- PERFORM RECOGNITION EXPERIMENTS - to quantify the improvement in recognition accuracy gained from using segment controlled deformation parameters.

Task 1:

The first task is the study of best match paths through dynamic programming matrices. Dynamic programming matrices contain speech sound similarity scores between time windows of two utterances where the rows and columns of the matrix represent the time in the known and unknown utterances being compared. It is generally observed in dynamic programming matrices that there are broad rectangular regions of good similarity scores connected by narrow paths of good scores.

The goal of the study is to quantify the deviation from linearity for narrow paths and broad rectangular regions.

As a result of this study we will be able to tell how flexible different types of segments can be. We may discover that utterances can be represented by a mixture of flexible and inflexible segments. The inflexible segments would be those that are adequately modeled as straight lines of slope 1. If this is the case, it might not only allow more accurate recognition results, it might permit more compact storage of utterance templates.

Task 2:

The second task of this research is the development of algorithms for representing and using variable template flexibility in dynamic programming pattern matching. In particular, we will investigate the use of A and B values in dynamic programming calculation using the following equation:

$$D_{ij} = S_{ij} + \text{MIN}\ (A*D_{i-1j}, B*D_{ij-1}, D_{i-1j-1})$$

Note that when A and B are larger than 1, this forces selection of $D_{i-1j-1}$ which represents movement along a path of slope 1. The proper values of A and B would be determined experimentally and carried along in templates with each time frame to tell the dynamic programming classifier which values of A and B it should use.

Task 3:

The first step in Task 3 is to generate templates that contain average parameters of several exemplars which are time-aligned with dynamic programming. The second step is to encode a deformation parameter expressing the allowable temporal variation for each segment.

Task 4:

The fourth task is the performance of recognition experiments. An attempt will be made to vary the basis vectors and similarity functions along with the constraints on segmental deformation in order to study their interaction.

## B. RECOGNITION OF SPEECH DEGRADED BY NOISE

The objective is the determination of an optimum method for reducing the deleterious effects of noise on the accuracy of automatic word recognition systems. Three alternate approaches will be studied.

Task 1:

The first approach involves an investigation of various recognition feature sets (basis vectors) to determine the feature set that provides the highest recognition accuracy under noisy conditions. The feature sets to be examined are:

- Linear predictive coefficients (LPC)
- Vocal tract area functions
- Autocorrelation coefficients
- Cepstral coefficients
- LPC derived Pseudo Formants

The first four feature sets are defined in the recent book by Markel and Grey.[1] The LPC derived pseudo formants are obtained from the LPC coefficients by setting the magnitude of the last coefficient to unity and solving for the pole frequencies of the resulting LPC transfer function.

To evaluation of these feature sets will determine the best recognition set for a wideband quiet environment, for a telephone bandwidth quiet environment, and for a noisy telephone bandwidth environment. The overall optimum feature set will then be selected. Incidentally, as a by-product of this evaluation, an optimum feature set for speaker independent recognition will also be determined.

Task 2:

The second approach involves the use of a noise cancelling filter applied at the input stage of ITT's Kernel recognition system. The method assumes a means for determining a noise signal $w_1(n)$ that is highly correlated with the actual additive noise signal $w(n)$ and uncorrelated with the speech signal $s(n)$. If $w_1(n)$ can be determined, then it can be shown[2] that an adaptive filter can be constructed whose output is a maximum likelihood estimate of $w(n)$, and the signal $z(n)$ is then a maximum likelihood estimate of the clean speech signal $s(n)$. Thus a proper selection of $w_1(n)$ will lead to a filter that effectively removes the additive noise component.

Various schemes for generating $w_1(n)$ will be considered in the study.

These schemes include:

- Setting $w_1(n)$ equal to the average background noise during periods when it is known that no speech is present;
- Setting $w_1(n)$ equal to the updated average signal during periods classified as silence;
- Setting $w_1(n)$ equal to the LPC residual error.

[1] J. Markel and A. Grey, "Linear Prediction of Speech," Springer Verlag, 1977.

[2] Widrow, et al. "Adoptive Noise Cancelling: Principles and Applications." Proceedings of the IEEE, Vol. 63. No. 12, December 1975.

Task 3:

The third task involves the investigation of a noise-reduced LPC parameter set recently proposed by Sambur.[3] This parameter set is determined by subtracting a term proportional to the residual signal power from the diagonal of the autocorrelation matrix used to determine the standard LPC set. The new parameters have been shown to provide a more accurate representation of the speech spectrum in a noisy environment. These parameters should provide a superior feature set for recognition purposes.

## C. SPEAKER INDEPENDENT RECOGNITION

In order to develop word recognition algorithms that will be insensitive to the characteristics of individual speakers, signal parameters will be investigated that carry very little information about the speaker. Three such parameter sets are LPC-derived pseudo formants, orthogonal LPC parameters and vocal tract area functions. LPC-derived pseudo formants are obtained from the LPC coefficients by setting the magnitude of the last coefficient to unity and solving for the pole frequencies of the resulting LPC transfer function. Assuming that the LPC-derived transfer function is:

$$H_n(z) = \frac{1}{1-a_1 z^{-1} - a_2 z^{-2} \ldots . a_n z^{-n}}$$

then the pseudo formants are obtained by setting $a_n$=1 and solving for the $a_i$ values and then finding the corresponding poles. The poles are now on the unit circle and have zero bandwidth. This result is a natural consequence of the fact that the last predictive coefficient is a product of all pole moduli of the vocal tract filter. By making the product unity, the individual pole modulus becomes unity, signifying that all poles are located on the unit circle. The pseudo formants are closely related to the actual formants, but unlike the formants, the pseudo formants vary smoothly across an analyzed utterance and can be easily labelled. Due to the normalization of the speaker sensitive bandwidth information, pseudo formants should be more effective than formants in providing a speaker independent representation of a speech word. The ability of pseudo formants to provide speaker independent recognition will be thoroughly examined and compared to other recognition features in ITTDCD's internal recognition system.

[3] To be presented at Acoustical Society of America's meeting in Miami in December, 1977.

A recent experimental study has shown that by an appropriate eigenvector analysis of the linear prediction parameters, a set of orthogonal parameters are obtained that can be used to achieve a high-quality synthesis of the original utterances. The interesting aspect of these orthogonal parameters is that only a small subset of the parameters demonstrate any significant variation across the analyzed utterance. The remaining orthogonal parameters are essentially constant and, for purposes of synthesis, are completely specified by their measured mean values across the utterance. In a later experimental study it was shown that these remaining orthogonal LPC parameters were associated with the speaker identity and characteristics of the channel. Thus it may be assumed that the orthogonal parameters with the most variation were conveying information about the identity of the spoken words and very little information about the speaker. The speaker independent recognition potential of the higher order orthogonal LPC parameters will be investigated.

Vocal tract area functions are another attractive set of speaker independent recognition features that will be examined. The advantages of the vocal tract area function for use as speaker independent recognition parameters are explained by the fact that the vocal track length can be estimated from analysis of LPC parameters, and then the vocal tract length can be normalized to a standard length.

### D. LARGE VOCABULARY RECOGNITION

The representation, storage and retrieval of speech reference data has been, and continues to be, a major problem in automatic speech recognition of large vocabularies. Computational limitations of general purpose computers have led to the emphasis of reference data coding in terms of rules of syntax and phonological rules. A compilation of phonological and syntactic rules is a monumental task that is, at present, far from complete. On the other hand, storage of reference data as single utterance template data is more easily implemented but more difficult to use because of the large amount of memory that must be searched. The solution proposed here is to make pure template information more useful through better information retrieval search strategies for large data bases.

The primary vehicle for speeding up the search process is to have several data bases of differing size and to allow the smaller ones, which are more quickly searched, to control the search of the larger ones. The question of what to put in the smaller dictionaries may be answered in many ways. It is suggested that compressed speech, with differing degrees of compression, be used to fill up the smaller dictionaries. Speech subunits, such as phonemes, may also be used to achieve a more compact data representation as well as well known parametric speech compression techniques. Storage of template data in nets and trees as used in the HARPY system at CMU will also be investigated.

### E. WORD SPOTTING AND CONTINUOUS SPEECH RECOGNITION

The recognition of continuous speech and the ability to "spot words" in a stream of continuous speech is not an unsolvable problem for carefully pronounced speech with a good signal-to-noise ratio. ITT is not planning to research the issues of semantic information processing nor syntactic analysis to support acoustic analysis. However, ITT believes that research into temporal variation and allophonic spectral variations will yield sufficient information to allow useful continuous speech recognition and word spotting systems to be built for cooperative speakers. The inherent temporal variability found in continuous speech utterance can be satisfactorily modeled by new techniques of dynamic programming. The inherent spectral variations caused by coarticulation can be satisfactorily modeled with principle components analysis. Dynamic programming research was mentioned above. As for analysis of spectral variation using the principle components technique, ITT plans to study this area and combine the results with dynamic programming results. A large amount of data will be analyzed to test the power of the resulting techniques.

### F. ISOLATED WORD RECOGNITION

The objective is to develop an isolated word recognition system that will serve as a kernel system for research in the other areas mentioned above. The specifications for this system are listed in Table 1. The goal is to implement the system on ITT's fast processor. Dynamic programming will be used in the classifier stage. Perhaps the most interesting aspect of this system is that it could potentially be implemented on a dozen chips costing less than $100 for parts if use is made of a single chip band pass filter bank.

## III. GOVERNMENT SPONSORED WORK

ITTDCD has gross revenues in excess of $60 million a year from Government contracts in the area of data and voice transmission and data handling. ITTDCD has contracts with the Navy and NSA for research and development in low bit rate and/or secure voice communication research and development.

TABLE 1

PERFORMANCE SPECIFICATIONS FOR ISOLATED WORD RECOGNITION SYSTEM

| | | |
|---|---|---|
| (1) | TYPE OF SPEECH | ISOLATED UTTERANCES, COOPERATIVE SPEAKER |
| (2) | VOCABULARY SIZE | UP TO 50 POLYSYLLABIC WORDS |
| (3) | ACCURACY | BETTER THAN 99.0% |
| (4) | TYPE OF SPEAKER | SYSTEM ADAPTATION REQUIRED: I.E. A TRAINING PERIOD IS REQUIRED - NO RESTRICTION ON USERS LANGUAGE OR DIALECT |
| (5) | RECOGNITION SPEED (USING EXISTING RAM PROCESSOR) | 10 SEC. MAX<br>1 SEC. DESIGN GOAL |
| (6) | SPEECH INPUT DEVICE | A TELEPHONE HANDSET (BUT NOT OVER DIAL UP PHONE LINE) |

## BIOGRAPHICAL SKETCH

### George M. White

George M. WHITE, Manager of San Diego Laboratory, ITT Defense Communications Division.

George M. White graduated magna cum laude in physics from Michigan State University in 1964. He received Danforth and Woodrow Wilson Fellowships for graduate studies at the University of Oregon and the University of California at Santa Cruz. He received his Ph.D. degree in theoretical chemistry from the University of Oregon in 1968. He spent two years, 1968 to 1970, at Stanford University Artificial Project on a National Institutes of Health post-doctoral study fellowship. In 1970, he joined the Xerox Palo Alto Research Center where for the next seven years he managed Xerox Corporation's automatic speech recognition project; designed many internally constructed recognition systems, some of which achieved notably high recognition scores for English; monitored speech processing technology for Xerox Corporation; advised Xerox venture capital groups on corporate acquisitions of speech technology firms. In 1977, he joined ITT Defense Communications Division where he is Manager of the ITTDCD West Coast office, which carries on research and development projects in speaker verification, speech recognition and speech compression.

Omit
TO
END

# WORKSHOP[1]

5

DR. EDWARD HUFF, CHAIRMAN

NASA AMES RESEARCH CENTER
MOFFETT FIELD, CALIFORNIA

## INTRODUCTORY REMARKS

As I indicated this morning, the conference can be characterized as having three distinct groups of representatives, and perhaps some in between: artificial intelligence researchers, technology users, and manufacturers. One of the challenges for this workshop is to bring together these related but somewhat disparate groups, each with their different ambitions, and see if we can get some kind of effective technical interaction and information exchange going.

Now, as I also indicated this morning, what we'd like to do first is allow an opportunity, probably lasting 15 to 20 minutes, or 40 minutes at the outside, for prior speakers, or people who did not feel that a particular topic close to their heart was brought out well enough, to make further comment at this time. After that period we will go into more of a structured set of issues which the participants and the audience can discuss.

It has come to our attention that a number of individuals representing different organizations would have liked to have participated formally in this conference, but for one reason or another were unable to get on the schedule. One of these is Dr. Bromley, from the Naval Ocean Systems Center, and so I thought we'd start off with him. He has a few comments to make. Please limit comments during this 15 or 20 minute period to four or five minutes.

Dr. Keith Bromley: Thank you. I'd like to take just two or three minutes of your time to tell you about some of what we think is rather revolutionary work that we're doing at the Naval Ocean Systems Center in building new types of devices, and new types of systems that I think are very strongly applicable to the voice technology area. I know nothing about voice technology, so I plead ignorance, but I am a semi-expert in optical signal processing. I'd just like to tell you a few words about what we're doing.

[1] Taped workshop presentations and discussions were edited during transcription with the avowed objective of preserving content, intent and style.

Looking at optical processing, it has two great strengths to offer in the signal processing world. The first strength is that you can multiply extremely rapidly with optics. Just the light passing through a transparency does an analog multiplication. The light intensity coming out is a product of the intensity going in times the intensity transmittance of the film. If you think about it, it takes light about a picosecond to pass through a piece of film, so you can do an analog multiplication in a picosecond. That's a pretty good number.

The second strength of optics is that it has a two dimensional nature. If you have an image shining on a transparency, and that transparency has a 1000 by 1000 resolution points, then looking at the light resolution coming out of the transparency you literally have a $1000^2$ products. Well, if you can perform $10^6$ multiplications in a picosecond, that's $10^{18}$ multiplications per second, you're doing some pretty darn good processing rates.

Now there's been a lot of hangups that have kept optical processors from achieving those kind of rates. The traditional bugaboo has been in getting the electrical-to-optical and optical-to-electrical conversions at the beginning and end. We have developed a system which we feel is not optimum because we're certainly not doing $10^{18}$ multiplications per second, but we are doing $10^{10}$ multiplications per second in a system which, when the first one is delivered in about a month, will be about the size of 2 cubic inches. It'll look like a device you can just wire into your circuit board, and if you didn't know there were optics inside of it, you'd never have need to know.

Basically our device consists of three components: a light emitting diode, a photographic reference mask, and a two dimensional charge coupled device array. The input signal is discreet in time but analog in amplitude, and is fed into the light emitting diode that converts the electrical signal to an optical pattern. This optical pattern then plugs the transparency, so that if the transparency has say, a 1000 by 1000 resolution points, you literally do $10^6$ analogue multiplications in the time that it takes light to pass through that transparency. Each of these products is then stored in the area array charge coupled device. Let's assume that it has 1000 by 1000 resolution cells also, although off the shelf right now it's typically 500 by 500, so you have $10^6$ products stored there. Now if you feed in one clocking pulse which shifts all of these products up by one resolution cell spacing, and you bring the next value into the light emitting diode, then it multiplies the mask, gets added to the charge packets that are already there in the CCD, and you can shift it up by one more cell. After requesting this process, by the time you've shifted all the way up the charge coupled device, you've kept track of all of the sums of all of the multiplications of your input signal by whatever function, or matrix, is stored on the reference transparency. Now, if the reference transparency is of the order of say 512 by 512, then you're simultaneously cross-correlating every single time window, a sliding time window of your input signal with 512 reference masks in parallel. Now these refernce masks could be coded to correspond to different words, and input signals could be time sample values of speech data, frequency values, LPC coefficients, or any vector that you wish to simultaneously cross-correlate with 512 references in parallel. That has to be a very, very useful device, especially when it can be reduced to something of the size of one chip.

Another thing this device could be used for is each column of your reference mask could be the various sign and cosign waveforms for different frequencies. Then the output of your charge couple device is purely a Fourier transform. Or it could be programmed to do a Hadamard-Walsh transform. Any linear transform you want to perform can be programmed into this particular mask. Now all this mask does is just a discreet Fourier transform. Now a discreet Fourier tranform takes N squared multiplications if your input signal is length N. And if a multiplication requires one unit of time, then it takes N squared units of time. Now an FFT reduced that to N log N multiplications, or N log N units of time. Now we're going back and doing the plain ordinary discreet Fourier transform in this device, but we're doing all N columns in parallel, so that means that even though we're doing N squared multiplications, it only requires N units of time. So this device is even faster than an FFT algorithm by the factor of log N.

I just had cause to do a comparison of our little device with a popular system that I've heard several people reference--the floating point system Model 120 B array transform procesor. I'm taking a look at the device that we should have ready in the coming fiscal year. Our device is something like 500 times faster than the FPF system. They do a 1024 point Fourier transform, in I think 17 miliseconds; we do it in 30 microseconds. Ours is one card instead of 3 feet of wrap space. Ours ultimately will cost under $1000 and theirs costs on the order of $80,000 to $100,000. Ours will use maybe 24 watts of power compared to 24 kilowatts of power, and the list goes on and on.

So, I thank you for letting me tell you in a brief few minutes that this technology is right around the corner. And maybe at the next meeting of this group I can actually have a working device that I can show to people doing these various operations. Thank you very much.

Dr. Edward Huff: Thank you Dr. Bromley. We could probably tolerate a question or two at this point if they are brief.

Dr. George Doddington: I don't have a question, but I have a comment. I appreciate your comparison of the speed of the optical processor with the AP 120 B, but I think you were unfair in not pointing out that the two devices are not at all the same. The array processor is in fact a very high speed general purpose computer.

Dr. Keith Bromley: A very good point. But if anyone wants copies of technical papers, just write to me or telephone me and I can send you a copy of the papers that we have.

Dr. Edward Huff: I don't know if he is interested but Dr. Rex Dixon, IBM, may have a few words for us.

Dr. Rex Dixon: For the people here who aren't aware of the work that has been going on inside IBM in speech recognition, I thought I would apprise them of the fact that we have been active in speech recognition since about 1955. For a period of about 4 years we poured about 50 manhours into the area, primarily discreet word recognition, back in those days. A number of papers were published, and then there was spotty work in the company up until 1967 when we started the first of a series of four one-year contracts with RADC in continuous speech recognition. The level of the work was about a two man-year effort, and during that four year period, a system configuration and preliminary performance data were collected. I'll make some comment about those in a moment.

The work that was done was specified, in terms of its size, by RADC. We were charged to use a 250 word vocabulary and to come up with a model --a language model--that would work in a continuous speech recognition framework, and would allow for flexible sentence level generation. We came up with a finite state model, 250-word system (this really doesn't mean anything but it gives you the idea of the kind of flexibility involved) that was capable of generating something on the order of 14 million sentences of 7, 8 and 9 words in length. On that language, we achieved recognition levels, at the sentence level, for one talker which were on the order of 40 percent. I think that was the best we ever did. Word recognition levels were close to 90 percent, and sometimes better than 90 percent, but it was, as I say, promising. The front-end performance that we were observing at the time was, we thought, also very promising. We were achieving segmentation accuracies on the order of, combined, less than 10 percent error; that is, combined in the sense of extra and missing segments. Classification accuracy in the phonetic domain, that is, in a classical phonetic sense, using four different talkers, ran from about 70 percent accuracy up to as high as 84 percent accuracy. The formal tests that were done at that time, as I said before, were very preliminary, but they were small data sets.

Then the work was taken out from under contract to Rome and moved to Yorktown Heights, New York, to our Thomas J. Watson Research Center, which is where I'm located now. It was manned up to a level of about 14 people at that time in 1972. It has continued since that time, and several systems have been built. One of the things that I think has characterized the work at IBM, in contrast to the work that has been reported here on the first day, is that the entire approach at the top end, or what you might call linguistic processing, has been on the basis of information theoretic, statistical decoding, modeling procedures. The people who have been doing the work come from the area of information theory, rather than from speech, and the systems that have been developed are functioning in simulation on a large general purpose computer, a 370 Model 168. We are looking into natural language. The particular natural language that we are looking into now is a very large body of laser patent texts, and we are using statistical modeling techniques in order to characterize the properties of the language, rather than using linguistic or grammatical structures. Actually, we do have some structures, or some models, which do things like associate a part of speech value to lexical items. We aren't using those in any regular way at this time, but all of the work that is being done is, in fact, in a statistical modeling information theoretic context. At any rate, there are two current systems. One system is a full, phonemically segmenting frontend, and it was just run, as a matter of fact, on the same language as was reported for the HARPY system at CMU, for one talker. The system did achieve 97 percent sentence accuracy with 2 decoding problems that were alleviated in one case by letting the system decode longer, and in the other case by raising a threshold. The one error sentence was one word confusion between "I am" and "I". So we're very pleased by the fact that this system is operating well, and we have another system which has a nonsegmenting frontend and operates at the 10 millisecond level. That is, it takes a segmental input 10 millisecond segments corresponding to phonemic events, with phonemic labels being the characterizer, and with that one, with this Rayleigh language I mentioned to you before, we're getting 94 percent accuracy at the sentence level, and better than 99 percent accuracy at the word level.

Dr. Edward Huff: Thank you. We can entertain a question or two for Dr. Dixon, if there are any? Is there anyone else in the audience, or at the table here, who would like to embellish on his prior comments or comment about areas that were not gone into in great enough depth earlier in the proceedings?

Dr. Wayne Lea: I'm not sure if this is the appropriate time. I'm Wayne Lea from Speech Communications Research Laboratory. I'd like to pose three ideas to this group and encourage consideration of these throughout work in speech recognition. One is that speech data bases be compiled someplace and made available, so that one group can, at least, act as a clearinghouse. Any time somebody wants to do something in the area of speech analysis--for example, the isolated word recognition test--they would then be able to obtain one of these data bases for their work rather than constructing it all over again. For example, I've heard data bases quoted of 20,000 isolated words with a number of talkers involved, and obviously it seems a shame if somebody has to do that over again to establish the effectiveness of their system with different speakers. That's also true with continuous speech. Now, the IEEE has a subcommittee on speech recognition, Dr. George White, TI, is the chairman of that committee, and in that committee there is a data base committee which is intended to address specifically this question of speech data bass. Jerry Wolf is, I guess, chairman of that subsubcommittee.

It seems to be a very important part of our activities, being chartered right now by the government to review this field, to do what we can to bring things together that can help everybody avoid duplication. So I would recommend to you that something be done in this area, and I suggest that if you're interested, that you would provide to me or Jerry Wolf all the speech data bases that you have that might be made available. Knowing the form the data are in is also very important. For example, there are a number of reported data bases that were compiled and put into a box so that only the results that came out of the front-end of a recognizer were preserved. I would recommend to each of you that you provide either analog or digital tapes, before they are put through a lot of processing, so that these might be made available to other researchers.

There also seems to be a complimentary need for summarizing all the applications that we've heard here in this workshop, and we, of course, know there are others elsewhere. I would recommend, and I have talked to Commander Curran particularly about this, that someone write a paper which presents in some reasonably succinct way an overview of all these government applications, what their approaches are, and in some sense provides to industry and to government a full picture of what's going on. I believe this is one of the intentions of the workshop and I whole-heartedly endorse it and would like to see it done promptly.

The first idea that I hope we will discuss at this workshop, is the feeling I have that there is a serious but interesting gap between the research and the applications, and I'd like to see more bridging done. I see very fascinating work in the field, and I see some good, hard evidence of the need for better techniques, but I have not seen previously as much interaction as we've had here in this workshop to bring the technology together with the researchers. I would like to see us do that more. I also have the feeling that there is a comparable gap between the continuous speech recognition (or speech understanding) systems area and the isolated word area. And it seems like something in the middle of that gap would be an appropriate area to be attacked by other workers in the field.

In addition I'd like to add one other thing, and that is that on the first day, when I gave my presentation, I did mention some recommendations about the future which we weren't able to look at, and if it seems appropriate at some time I'd be glad to take a look again at those.

Dr. Edward Huff: Well, actually, that's a good lead-in. I'll tell you what happened. We got together early this morning and tried to put together various questions or issues that we culled out of the three days, and also to put them together the inputs that were received from the audience. As a matter of fact that's being further integrated right now as you're watching because we just received some new suggestions. Of the five inputs from the audience, two of them fit into the structured issues that you just mentioned. Three are different from what we had considered, although they're quite familiar issues now that we look at them. So, the only problem is really where to begin, and I think you've done that for us. I think we should begin at the topic of data bases. That was item nine in our list, but it's just as good as anyplace to start. Do we have any discussion or comments about the utility of data bases, where they exist, or who should coordinate them? Dr. Lea made some suggestions. Are there any liabilities to data bases that we can see? Would it hurt anybody? Would it prove to be perhaps misleading, or ineffective in some way, or are there only advantages to be derived?

Mr. Michael W. Grady: Let me just add one thing to that. For our discussion I think there are two kinds of data bases we might keep in mind. I'm interested in a large number of training applications and military applications that are required, and getting analog tapes of transmissions and typical noise environments that occur in real life. These kinds of analogue recordings are as important as the nice, clear acoustically perfect kinds of data bases that I would imagine the folks like Wayne Lea are interested in. So keep that in mind in the discussion.

Dr. Edward Huff: Does anyone have data bases available?

Dr. Jared J. Wolf: Yes, could I say something about that? Wayne alluded to a rather modest effort that the IEEE technical sub-subcommittee is trying to do something about. We didn't feel that we had the resources or the desire to act as a clearinghouse for data bases themselves, or be a central repository. But we figured that it couldn't hurt at least to get some information around about data bases that do exist and that may be available under some circumstances. And so at a couple of technical meetings I've tried to promote this and haven't been very successful. What we've done is thought about it for about a half an hour and put together a little questionnaire, kind of a standard format for describing a data base, and we're forming it, if you will, as a data base data base. If we can get people to volunteer to describe their data base in this relatively standard way, then I'm simply acting as a clearinghouse. We're going to collect these together and make them available, and there should be a couple of announcements coming out in the professional journals in the next couple of months. I put these in a few months ago and the publication lead time necessitated a delay of a few months until next spring. But the idea is that this will be available to the people who have data bases and also to anybody who might be interested in finding out what might be available. We are finessing the question of who pays for the labor of duplicating your digital mag tapes or your analogue tapes. What we're trying to do is to put together the have and have-not groups and let them work in out themselves. The essential idea is getting a little bit of information about what is available and determining whether it can be used.

So, our effort is the following: If you want to find out about something, write to me (I've been immortalized in the roster here, so you all have my address) if you have something that you would like to make available, or under some circumstances you could conceive making available to somebody else. If you write to me I'd be happy to send you a questionnaire and then we'd at least get your data base described in the same format that the others are described in. But you've got to take the initiative to write to me, or look at my announcement in the journals, so that I can send you the form. I'd rather that you not take the trouble of writing me a letter and then leave out four or five essential aspects of the data base. The questionnaire is designed to get you to say a few standard things about it.

Dr. Edward Huff: Are you planning to distribute the questionnaire to all of the participants at this conference by any chance?

Dr. Jared J. Wolf: I wasn't planning on it, but I could do that. Count on it.

Mr. Marvin B. Herscher: Jerry, you might, before you distribute it, be sure that you have questions such as what type microphone is used; it's characteristics, bandwidth, background noise, and whether there's pre-emphasis in the system. These factors all make tremendous differences in terms of how valuable the data are to anyone other than the person that collected them.

Dr. Jared J. Wolf: Good point, thank you.

Mr. Thomas B. Martin: Really, I feel that this is the responsibility of the government. It was tried by the speech processing committee under IEEE sponsorship 7 or 8 years ago and it got nowhere simply because there weren't any funds available and because there weren't enough volunteers and time to do it. In the end there may have to be some duplication. Also, there are no standard for recordings, which means that finding exactly what conditions were recorded under, particularly in the digital domain, can be completely different from disc to disc. The government has funded sufficient work up to this time so that very large data bases exist. For example, in the 60's the Post Office, I think, acquired tape recordings of roughly 500 of their employees on both isolated digits and connected digits in both quiet and in 88 decibels background noise. That data base has been available to anybody who wanted it for almost 10 years and nobody has ever wanted it. Moreover, there is no depository; those tapes are now scattered around different placaes, probably 2 or 3 places. There's no depository and no sponsorship, and it won't happen spontaneously within industry.

Dr. Edward Huff: Any further comments concerning data bases?

Mr. Sam S. Viglione: Yes, I think along the same lines as Mr. Martin, but nonetheless the feeling I have about the data base idea is probably not shared in the speech community. For example, years ago we were talking about establishing data bases for images that we had to try to classify, so that might have some consistency across the different concepts that were developing. The justification for that was very simple, it was tough to fly over and pick up enough foreign material, e.g., Russia, or pick up some missile site information. On the other hand, in speech we have a proliferation of data. It's very easy to come in and stop everybody wandering by the hall and say: "come on into my lab and speak to my computer for a little bit, I want to collect some data". The nice part about it, however, is that you can regulate how the data's collected. You know all the consistencies and inconsistencies of the data. You know the recording properties and you're putting it immediately in the format you want. I find it very easy to generate a data base for whatever type of data that we need for a specific experiment we want to conduct. In some cases it's easier for me to do that than it is to try to understand the format and the collection procedure that somebody else used, so that I can use his data effectively. So, I think that we have to be a little cautious in establishing data bases. I'm not sure whether it's cost effective for us to do it in some cases.

Dr. Wayne Lea: May I reply? It seems to me that this is a good point. There are times when clearly somebody's data base is of use to nobody else. We expect that. I don't think that disallows the case when they are of use. And one of the uses that I see has been nicely summarized in a recent paper in the IEEE transactions on audio (the speech-related transactions) in which they try to deal with the question of how to decide whether two computer techniques, or two devices for speech recognition, are working as well as each other. Often you find that one of the reasons why you can't decide which one is better is because you didn't try them on the same data. And it would be excellent I think, for IBM to have used not only the same language, but to have also taken exactly the same data as Carnegie-Mellon had, and to have put that through their system and said, "See, we did as well, or better, or less or something". I think that this is an excellent way to make us confident, and for the Government to judge relative performance. That's one way of standardizing one part of the problem involved in speech recognition.

Mr. Michael Nye: I'm Michael Nye and I just wanted to comment on that business. In sitting through the presentations during the week people talked about many things including a $299 device that was quoted at a medium recognition range of 98 percent. I don't know why you would call that medium, because 98 percent's pretty good to me for a device that would cost under $300 considering that devices up to $100,000 recognition systems produce the same kind of recognition accuracy with the same size or type of vocabulary. The point I'm trying to make is how do we know what accuracy really means? If you really want to do something meaningful with data bases, why not establish a criteria for a particular vocabulary. Make it a 16 word vocabulary, or just the digits, and decide on the constraints, and then whenever somebody wants to produce evidence of the great and wonderful things they've been doing, they can refer to that specific vocabulary and say, "this is what we've done". And if somebody else wants to quote accuracy rates, they're quoting a common standard. But we are getting to where I feel like I'm in a used car lot. I'm looking at a half a dozen different cars, which all appear to do the same thing, but they don't. I may not have expressed myself the right way but clearly my heart's in the right place.

Dr. Mark F. Medress: I'd just like to offer a little perspective on this data base problem. I personally feel it's very advantageous to be able to share data. In the ARPA Speech Understanding program there were initially five different people building systems, all of which were designed towards the same performance goals. We had a coordinating committee that was trying to establish procedures for sharing data, and we spent a lot of time meeting and trying to agree on formats and recording conditions. Some of you here, I'm sure, remember all those discussions. We basically never succeeded in finding a common set of procedures that we could follow, because there were just enough differences in the way people expected the data to be collected, and in the vocabularies and in the tasks, and so on. So, I'm all for our ability to share data, in fact we're involved in a program right now where we share data with other people to make those comparisons. It really requires a dedicated effort, however, and I think we have to be realistic about that.

Dr. Edward Huff: I think I could summarize by saying that we do plan on having additional conferences and meetings, perhaps not as large as this one, in the applications area. I know that this will be a continuing thing for us, those of us involved in applications, so data base standardization is an issue that will come up and we would like to work with those of you that have expressed an interest in it. We will certainly take it under advisement and try and do the best we can.

Dr. Robert Breaux: I'd like to expand on what Mr. Herscher had to say. That is, when we indicate what the data bases are all about, we have to bring in the psychological or human factors aspects as well, e.g., under what conditions were the data collected? In this way perhaps we can look at difficult vs. easy kinds of tasks, and so forth. It has been pointed out so well earlier, and most of us know that these kinds of situations are very significant in recognition.

Mr. Marvin B. Herscher: One further expansion on what I said yesterday is this. When you collect data, for example, in isolated word recognition, and there are very long pauses between words, and the speaker is very relaxed, that's quite a different data base from when you're out in real world trying to get data into the system rapidly and accurately. There are many different stress factors, both mental and physical, that are involved, and how a system performs under relaxed conditions is quite different from how it's going to perform when it's really being exercised, and those data base changes are quite significant.

Dr. Edward Huff: Thank you. There's just a panoply of interesting topics to get into. One that came up during various parts of the meeting has been references to better "front ends". At this point I'd like to put it out for discussion; is there a consensus as to the need for better hardware at the frontend of the recognition system? If so what is it? Are we going to put all of our hope in the advanced technology that has been presented, or are there other things yet downstream?

Dr. George Doddington: Okay, when you say better frontend processing, I think what most people were talking about was not better hardware, but rather better classification of acoustic data.

Mr. Michael Grady: Isn't that done by hardware?

Dr. George Doddington: No, no, it's not done by hardware.

Dr. Edward Huff: So the frontend includes the classification. Any comments on that?

Mr. Marvin B. Herscher: I think, George, that what you really mean, is better descriptions of acoustic events. You don't mean the final classification, you mean the classification of acoustic events, in terms of...

Dr. George Doddington: Yes, that's what I meant to say.

Mr. Marvin B. Herscher: As opposed to recognition decisions...

Dr. George Doddington: I wanted to respond, to head off the false impression that people might have from what you said. You mentioned the term "hardware" and I wanted to clarify, or at least open the argument about whether or not the better frontend means better signal processing in terms of Fourier analysis, or filter banks, or pitch tracking, etc., vs. acoustic classification (what sort of phonetic events are occurring), and I believe most people who said we need a better front end meant that we need better classification of acoustic phenomena, and not a better signal procesor.

Dr. Edward Huff: I'm glad you brought that up, because the next item we thought to expose as an issue was the desirability of phonetic analysis as opposed to other forms of classifications. Is there a concensus at this point as to the need for phonetic analysis specifically? There are different approaches to this apparently, and of course that gets back to the issue of what we mean by the categorization or classification that takes place at the frontend. So those of you expert in this business can comment perhaps.

Dr. Mark F. Medress: I'll say just a word. Phonetic classification has served as the backbone of our system development for continuous speech understanding, and it's something that we feel fairly strongly committed to. I think that the developments in this area are very exciting. We're hearing about some new ideas here, and within the last year, for ways to take advantage of acoustic information without requiring some specific phonetic analysis. I think there are a lot of arguments that promote phonetic analysis as a framework for introducing linguistic regularities, phonological regularities, acoustic phonetic rules, information about language and its manifestations in the acoustic waveform, can be used to do a better job of describing events. So that's kind of the perspective that I have. And yet, at the same time, I think we're seeing other approaches for capturing this information, for exploiting those structures and things like networks and that kind of thing. I'd really like to hear some discussion about how people feel about this general topic.

Dr. Robert Breaux: I'd like to point out that I wonder often why the ARPA work repeated the development of hardware parts of the frontend that some of the manufacturers may already have had. They seem to be so adequate, so accurate, and to provide such good information and such good sound classification as opposed to phonemic kinds of information. I wonder is there a difference between sound classifiers and phonemic kinds of frontend? And, do we really need to couch this frontend analysis in terms of phonemes, or can we assume that the machine that classifies sound simply has a different ear than the human, and that regardless of the fact that it's a different ear, that it's a reliable ear, and that we don't need to look at it in terms of some structure that we want to impose on it? Suppose we allow it to impose its own structure on that signal? To me, that's similar to what the isolated word recognition systems are doing.

Dr. Jared J. Wolf: I believe that Mark just stated the case for phonetic or phonemic level classification very well, and I think he already gave the answer to your first question.

Dr. Edward Huff: Apparently he missed it.

Dr. Jared J. Wolf: A phonetic classification is the basis we have for understanding speech and relating linguistic phenomena to actual speech acts. These are, particularly, things like co-articulation rules, chronological rules, and regularities that we use to describe language and that we wish to use to recognize them. Something that operates purely in terms of sound classes, I believe, tends to cut itself off from that sort of description. And I believe that's the answer.

Now, as to the question of whether we need it or not, that I think is still an open question. I think we originally worked from an assumption that we did need it. It was one of Newell's dogmas of speech understanding that we need to apply knowledge to the speech everywhere we could, and this looked like a good way to do it. I don't think it's been proven, however, and it's very much an open question right now. As I wanted to say the other day, and I didn't get a chance to, I don't think that the very glamorous success numbers that come out of the ARPA project have necessarily discredited the a-phonetic-phonemic classification concept. I don't think we know how far that can be pushed right yet. I'm sure we're going to find out a lot more about that in the next few years, but I don't think we know right now.

Dr. Edward Huff: I have the feeling that what we've called sound classifiers are probably tapping some of the redundancy that is contained in those more formal ways of looking at things. But, the matter must be looked at that to find out just how that's true.

Dr. Jared J. Wolf: I'm sure. It is clear that they are tapping a redundancy. But I think it is not a reliable way of considering the matter. I don't believe so, because I don't believe the speech producing mechanism is a reliable producing mechanism. That's one hope that segmental analysis has; it's going to try to take into account some of the inconsistencies of speech and classify those things together, just in the way that a human does. A word initial "P" and a word final "P" are not the same, but they are classified the same by a human, and we're trying to imitate some of that unifying in a machine.

Mr. Michael Grady: Some of the work that we did, that I reported yesterday, in our final initial attempts to look at isolated word speech recognition, was a success only because we found that that sound classification approach was incredibly reliable. Over time, over many utterances, it extracted something, I don't know what, out of the speech that was always there.

Dr. Wayne Lea: I'd like to say that I started the ARPA project pretty much from the same viewpoint that everybody else in the project had: that phonetic analysis was necessary, that a linguistic model was the way to go in speech recognition. I think I've kind of changed in a way, at least I think that I've learned from the people who take a mathematical approach. There's no question in my mind that phonetics is a proper part of speech understanding in the long run, because, in fact, it involves representing classifying the human in both perception and in production distinctions, some of which are important, and some of which are not important. And, of course, it's the ones that are important that we're trying to capture when we do speech recognition. But I think that you have to admit that the classifier-like systems--the systems that use mathematical processes, discriminate functions and various other mathematical schemes as Newell called them--are generalized input-output techniques and are much more rapidly getting to that answer. It's very clear because they are looking for cues that are useful to the machine immediately, and you can actually find cases where you train the machine with one example, and by a mathematical scheme it will do very nicely with that same speaker as long as that speaker is not markedly changing his speech. But that is exactly the point: when you get to extended speech, and when you are dealing with extended time periods, and an extended speaker population, that's when you're getting into asking: "What is it that's common among all these instances of speech?" And that is the phonetic structure, and of course, we will ultimately want it. Now, in HARPY, for example, you can't find it, primarily because it was very carefully fine tuned into the structure of the system. And if you can say anything, the machine knew a great deal of speech, but it wasn't conveyed to the user, or to the builder of that machine. In fact, right now I understand that Bruce Lowry is trying to analyze that system and understand exactly what it was phonetically that seems to be working so well.

I think that the summary of all this is that in the immediate future the mathematical schemes have a real role to play, and they really will capture quickly the success that's needed in an application. But in the long haul, where we're trying to deal with the more general problems, we're going to have to bring in and constantly be aware of the phonetic regularities that are there.

Dr. Edward Huff: Any further comments? One of the issues that we thought about last night is the adaptiveness of the system. On the one hand, we have speaker independent systems, or the concept of the speaker independent system in which it is implied that some regularity is known about the population and presumably in the ideal case very little more needs to be learned. On the other hand, we have the speaker dependent systems requiring instant or immediate training of differing degrees. Somewhere in the middle, perhaps, is the concept of dynamically adaptive systems. Indeed, a few speakers have discussed the idea of the system tracking along, so to speak, either to learn better what it should have learned in the first place, or to keep track of periodicities in vocalization, or the environment that is perhaps changing behind the individual. Do we see a need for adaptive systems in this sense, or perhaps some other sense? If so, how will they come about? Is anyone specifically working in that area?

Mr. Michael Grady: Yes, I think what we can address is in training systems, in particular, where we really see that adaptation is an important thing; particularly in the sense of having the system move along with the student or trainee during the course of his progress. Take us as a good example. When I was first initiated to the GCA tasks I had a stutter. Is it glide "slope" or glide "path"? And, is "turn right heading zero two six", said like that? Once the student becomes familiar with what he's doing and he's got it down cold (and, in fact, he thinks he's pretty hot) it's "turn right heading zero two six on glide path!" But that kind of speaker adaptation is really different than the question of whether or not a speaker--or a system--should or should not be trainable. The notion of speaker training, or machine training, as I brought up in my discussion yesterday, is not yet without significant drawbacks. Now, how to do that speaker or system adaptation represents some of the kinds of questions that people who are more familiar with the nits and grits of the speech process itself can help us out on.

Dr. Edward Huff: I can't help but comment, Dr. Doddington, that there must be something in that long term learning curve, that I believe you presented, having to do with reduced variability in vocalization over time. Something's happening to the process in your application, is it not?

Dr. George Doddington: I was going to say something about adaptation this morning, and I had to cut it out because I was running overtime. As far as adaptation is concerned with voice authentication, there are three motivations. One is to eliminate the bias in the reference data based upon peculiarities in a single session. Number two, and most important, is to track the adaptation of the user to the system in the short run. We saw in the case of the voice authentication that it wasn't all that short, however, being about 2000 sessions. And third, and I would say a distant third, is to track long term changes in a person's voice. I just don't think that's a very important factor.

Now we've tried adaptation in word recognition systems and that's a much trickier matter because you're not faced with a binary choice anymore, the probability of error is a little bit higher, and the disastrous effects of adapting at the wrong time is a tricky thing to handle. I would like to make a suggestion, though. I think that word recognition system builders might be able to improve the performance of the system by going through, for the first 15 days of usage, one pass of the vocabulary. If that's possible, they should incorporate in some way, an adaptation or updating of the reference file for the first and last usages. In other words, they should not use a sophisticated adaptation, not knowing the answer, but rather a very simpleminded, quick adaptation during the first sessions.

Mr. Robert Osborne: It seems to be that there are many applications where it is in fact difficult if not impossible to train. Dealing with the general public as we do over telephone lines, it's completely impossible.

Mr. Michael Grady: Oh, yes, I might point out that I definitely only meant the tactical and training applications.

Mr. David R. Hadden, Jr: What about stability considerations and that kind of thing?

Dr. Edward Huff: What kind of stability?

Mr. David R. Hadden, Jr: In the sense of adapting to the point where you can't tell two words apart. Is that a danger?

Dr. George Doddington: If that was addressed to me, all I said is that adaptation is a tricky process, and I was suggested that you might be able to get some of the benefits of adaptation without the trickery by going to an explicit known training class, for the first end sessions of usage. That's all I was saying.

Mr. Michael Grady: Isn't that what you did, Robert?

Mr. Marvin B. Herscher: I think I commented yesterday about adaptation. I think it's extremely task-oriented in terms of what you're trying to do, what you're trying to accomplish. It does, in most cases, involve some kind of feedback to the user to indicate that he's made a mistake. If, as I indicated yesterday, the user is busy doing something else, or isn't watching, or isn't alerted to the fact that a mistake has been made, then the system can very rapidly adapt to the wrong answer. And you can, considering the starting accuracy and the problems that you have, just as quickly diverage and get bad data, as you can converge. So, it's very much a function of what you're trying to do, and it can be very dangerous.

Mr. Leon A. Ferber: I just wanted to comment that it could be very dangerous, especially if you said a word and then you corrected it, and that word was mistaken, not necessarily because the person said a new version of that word because of some kind of a noise, or whatever. The person has to somehow observe what he said before, before you put it into the reference. It's very dependent on the application.

Another comment that I wanted to make is that from hearing people talk about adaptation and training, that we use all these words loosely, I mean we really don't always mean speaker adaptation. What does that mean? Does the speaker adapt to the machine? I just wanted to comment that there is nice work going on at BBN concerning the speaker adapting to a speech recognition system where the speaker himself, as opposed to speech recognizer, was fixed. And, on the other hand, you could have the machine itself learn, or adapt, and that's also quite loose. What's the difference between adaptation and teaching the machine? Where do you put the distinction? Does it depend on feedback to the user, or is it automated so that the user doesn't have to get involved at all? How much time do you invest doing that? For example, if you just want to enter three numbers by phone, are you going to spend half an hour training, or even ten numbers to train, and then enter three numbers? So it's very, very dependent on the application, mainly.

Mr. Marvin B. Herscher: This is just as a commentary to some of the things George was saying about improvements as you adapt in verification. His third part of the curve, where he interprets it as the speaker adapting to the system, which is probably correct, brings to mind something I probably should have mentioned yesterday, and Rex reminded me of today. Our experience with speech recognition systems in the field is that, although one would like to think that accuracy is very high instantaneously, that that isn't normally the case, except for some talkers. In general, we found many cases where accuracy continually improves, and where we were still going down in error rate over a period of six or seven months, at times. You just haven't reached an asymptote. Very little training was involved in terms of retraining retraining the words and speaking them over again, using the equivalent or same reference set, but the talker is learning how to make the system respond to him or her, and accuracy is very definitely getting better and better for the same reference pattern.

Mr. David R. Hadden, Jr: The question of adaption in a situation where you don't have time to train the system, the situation of being able to come on sort of wide band and lock on to somebody talking to the system, would seem to me to have a lot of appeal--if you could somehow control that. From personal experience, for example, I could picture somebody coming up to me and talking with some sort of strange accent, and then there's a few seconds where I'm storing and trying to figure out what the distortions are. And that might be a method of handling the problem where you can't train the system or the talker.

Dr. Edward Huff: Someone in the audience seems to have gotten interested in saying something.

Mr. Eugene Levin: It seems to me that one of the key elements concerning whether we are going to train the speaker or train the system is exactly this matter of feedback that Mr. Ferber talked about. If you initially have a fixed vocabulary, some kind of an averaged data base from a population, and no opportunity to customize it to an individual speaker, one method of adapting either the speaker or the data base is to provide feedback of what the computer recognized to the speaker at the time he states a word. Either it was recognized correctly, it was rejected as not being recognized as being in the data base, or it was recognized incorrectly. If it is recognized incorrectly, the speaker has the opportunity to ask for the next closest fit, if any. You have your choice, because one of them is going to get trained at this point: the computer can record its version of the sound that's being made, if it's one of the legitimate words in the data base, or the speaker can attempt, subsequently at other trials, if he gets rejected, to get a positive response from the computer for the word he's trying to state. So in either case, it seems to me that whether we're going to train the speaker, or adapt the data base, or gather more information on a larger population of speakers on how they say words, and how they think they say words, one of the key elements is to provide to the speaker--at the time he is talking--a response back from the computer of what it thought he said, before he confirms and sends. This would also take care of Mr. Hadden's problem, of having no training time, of having to come on the air and attempt to communicate. It gives the speaker the opportunity to confirm what the computer thought he said prior to send and the next confirmation.

Dr. Edward Huff: I think we might take a few more questions or comments from the audience.

Mr. Michael Nye: I have one comment and then a question for George Doddington. It seems, based on my application experience, that adaptive training would be really most appropriate for speaker independent systems. That is, suppose you get a talker from Milwaukee who wants to enter digits with a Milwaukee accent which might be slightly different, only slightly different, from the way somebody in Washington, D.C. might say the digits? I can see the need, obviously, for some adaptability there.

The question, George, concerns your offhand comment, which to me sounded very significant. That is, you quoted three reasons why you would want to have adaptability. The third reason is where you made your offhand comment. I got the impression that based on all these 1000's and 1000's of experiences in your test operation, George, that speech data really don't change much of the time.

Dr. George Doddington: No. You have to bear in mind that's a gut feeling.

Mr. Michael Nye: But if that's true, then for a speaker dependent system, adaptability shouldn't be important, should it?

Dr. George Doddington: There are three reasons for adaptability, and I think the least important is long term changes in the speaker's voice.

Mr. Leon Ferber: That's right. He says the conditions are changing. If the conditions are held exactly the same, then I guess that's what you mean by the speech data wouldn't change, under exactly the same conditions, right?

Dr. George Doddington: One of the conditions, several of the conditions, are the experience of the user, and other things. One of the conditions which I think is least important is the age of the user. The age of the user is how I characterize long term changes in the speaker's voice.

Dr. Donald Connolly: I just have a few observations that confirm a couple of the things that George Doddington said. One is that I found that where my speakers tried to modify their vocalization to make the machine happy, it messed up every time. The other is this. I believe that my kind of gut feeling for his secondary adaptation in the curve that he showed us is familiarity breeds contempt. It also breeds a certain amount of comfort. And I found the same kind of general effect with my people over time. When they finally felt comfortable with the beast, then things moved up. Thank you.

Dr. Jared Wolf: I have a couple of comments, based on past comments. There have been at least a couple of papers on how somebody's voice changes. Within the past four or five years, there was a paper by Enders, et al., in JASA that showed rather long term variation of up to twenty years, and that's perhaps I believe. And there was also another very interesting series of three papers in Electronic Communications in Japan, I think between 1972 to 1974. I can find them for you if you're interested. In one of them, the author attempted to look at how voice changed in a speaker verification situation, not speech recognition, and I don't believe he was very successful in tracking long term changes. But it seemed very clear from his experiments that there was at least a short term, random variation, and by short term I mean several months, and that you needed to be able to characterize somebody's speech over intervals of at least that long before you were doing a good job of tracking for purposes of speaker verification. I would recommend both these papers on that particular topic.

Dr. Connolly, you said that when somebody tried to make the machine happy, he invariably messed up. Does that mean that the machine really didn't recognize him? I presume that he was trying to speak very carefully in hopes that the machine would recognize him, right? Bob Rowan Clatt found the same thing in the LISPER project back in the late 60's, but in the thesis that Leon Furber referred to, the thesis by John McCool in which he was studying speaker adaptation to the machine, it was found that when the speaker received adequate feedback as to what the machine was misapprehending, if you will, he was able to modify his behavior very successfully.

Dr. Donald Connolly: I had the amusing experience last summer of being able to teach 3 RAF officers to speak with an Irish brogue in 15 seconds. Tell you about that later.

Mr. Gokal Gupta: I'm Gokal Gupta from BNR. I had a question for George Doddington. I just wanted to check this. As far as this user adaptability is concerned, it could be both audible as well as visual, you know. For example, in some of the applications that you showed, a person comes to the door -- he doesn't know whether the machine is asking him to say something or not -- it will say, "Howdy," and this and that. My question is, did you try the visual aid also? In other words, prompting by visual aid, as well as saying what message is flashing on the thing? Did you try to evaluate that aspect also?

Dr. George Doddington: We have used visual prompting, and we have used audio prompting. We have not made an evaluation of the different effects they have. An offhand comment I have is that the visual prompting is a little faster. I have a feeling, which I'm sorry is not verified by any data, that the audio prompting helps stabilize the user's voice. You may have gotten some feeling for that by listening to the recordings -- the response to the voice prompting is generally quite similar in timing to the prompting itself.

Dr. Edward Huff: I'm sure that we may very well get back to the adaptive aspects of the system. Although there have been a number of comments, and

I think quite a few answers, I can sense that there is a substantial consensus here. Therefore, I would like to switch now to the applications area, and put forward the question as to what we can discern about the requirements for continuous speech, as opposed to discontinuous speech recognition systems. The artificial intelligence work that's been done is highly significant, but at the moment, at least it has not processed in real time. There are "limited continuous systems" that have been discussed which are more real time, and some of them are right around the corner. But what are their requirements? How can we identify when a continuous system is needed vs when we can get along with a sound classifier? When do we know, how do we know, that a continuous system is more than just nice to have, and that it is really necessary? What are the limitations of sound classifiers, perhaps, that lead us to the conclusion that we need a continuous system?

Mr. Thomas B. Martin: There's really only one reason that determines whether you need an isolated system or a connected system, and that's simply speed. There is no other difference between the two. Furthermore, with an isolated word system, you can handle phrases. For example, let's take the ARPA project. "Tell me something about China" can be processed, as I said it, with an isolated word system, in that "tell me something about" can be treated as one phrase, and "China" can be one of the 200 or so options that ARPA had at that node. That happened to be their max number, and some of the things ARPA demonstrates can be done very simply with a real time-isolated phrase recognition system.

Dr. Edward Huff: Well, it's comments like that which lead us to ask the question I think.

Mr. Thomas B. Martin: Well, let me give you a task, for instance, that has to be done with connected speech -- digits at 300 a minute. You can't do that any other way.

Dr. Robert Breaux: I disagree that that's the only reason, and the problem that I have is called "user acceptance". We do not have uncooperative speakers, we don't have hostile people, necessarily, but we have people who are maybe on a short term exposure to the system. There's not really time for them to adjust to an isolated word system, particularly in the GCA application. The controllers are there for a week, they have just spent two to four weeks in various other applications of Tower Cab, ASR approaches and other kinds of talking, in which everything flows very naturally without isolated words. They're in this system for one week, and they're under pressure as naive trainees to be like their instructors, who are very smooth and fast and neat and sophisticated. And the instructors are somewhat against it as well. So speed is not the only reason. Speed may be the only reason in a very cooperative situation, but in a not-so-cooperative situation as I've described, there's also the matter of user acceptance, and for me that's a very important thing.

Dr. George Doddington: I made a wild comment in my presentation this morning which got vigorous negative response from the audience, and that is that in isolated word recognition the performance in terms of pure substitution rate is almost becoming secondary compared to other factors, such as the violation of the isolated word constraint. Tom Martin is much more knowledgeable about this than I am, and I'd like to get his comment.

Mr. Thomas B. Martin: I think, again, that you have to say something specific to a task. Even though I've talked for years to speech recognition systems, I don't consider myself really that experienced, because I have seen people that talk 8 hours a day, 40 hours a week, and I can't believe what they do. It may take months, and it may take weeks or days, but it becomes second nature to them, and the 10th of a second, or two-tenths or so that they can insert between utterances is natural to them by that time. I grant you that in one week you can't do something like that, and even though researchers think they talk a lot to machines, they don't do it for a living, and you've got to see some of those people to believe it. And so if you say you can't go 100 words a minute with an isolated word system, you can!

Dr. Mark Medress: I have a suggestion for a way of characterizing this problem. It's like a gut feeling, and it's only gut feeling, but if your task involved talking to people and talking to a machine, I would predict that it would be difficult to switch back and forth between pausing for one or two hundred milliseconds between words when talking to a machine, and then doing what you would normally do in talking with people. And I think, in particular, of what Dr. Connolly's doing with air traffic controllers. So I'm very interested in the results that you'll get soon from your experiments, and I think that's another dimension to the problem. Dr. Breaux suggested that one aspect is the amount of time that a person has to become familiar with the system, and Tom said that when a person is using a system continuously as his main function in performing his task, that he does adapt to it very effectively. But I think there are a wide variety of applications for speech input that want to use the sort of natural linguistic confidence we're all born with, for communicating verbally, in very much the same way as we would communicate with other people.

Dr. Donald Connolly: Mark is right. I hope that we'll get some good quantitative information on this in the next few months, where the user of the word recognition system has a number of other things to do. He is not going to make a living talking to a machine. I had some loaded experience in this, by being unwillingly sort of dragged into show business. During one period of three days last summer I was shifted between talking to my machine and talking to some fraction of a quarter million people that passed in front of it. My voice only lasted 3 days. I found that the first day it took me about 20 to 30 minutes to adapt to shifting modes of speech between what I used in talking to the general public for answering questions, and that which the system required to function with some degree of credibility. The second day it took about 20 minutes, the third day it took somewhat less than 10. I have a feeling this is something that even somebody in an even more complex intellectual task than briefing the public, air traffic controllers to wit, can do it, too.

Mr. Marvin Herscher: One of the things that bothered us when we first started working in isolated recognition was the important question whether people would be able to talk in isolation, whether it would annoy them, whether they'd be able to get used to it and be successful. I think we've had more experience than anyone else in the world, and the answer is, amazingly, that the average everyday person on the production line, the person that really isn't very well motivated, adapts amazingly well to speaking in isolation. You have to experience it to believe it. It's a very interesting experience, because it does take place, it does happen, they have no problem, they can turn around and talk to somebody else behind, and go back and talk to the machine in exactly the way they were accustomed to talking.

Mr. Michael Grady: I think we're running into a problem here similar to that which we ran into when we talked about adaptation. It's very difficult to talk about the requirements of when is continuous speech necessary, and when is an isolated word system sufficient, without speaking within the context of a very specific application. The question really is, what percentage of applications of everything in the world require connected speech vs what percentage would be adequate for isolated word recognition. You are bringing up topics here and subjects there where, yes, it would be very nice. There are definitely areas in which connected speech is necessary and would be extremely desirable, but on the other hand, if you simply took it from an economic standpoint, you'd ask how many industrial applications, how many users, how many millions of people could speak into an isolated recognition system. Let's make the assumption that isolated recognition might be less expensive than connected speech, because otherwise there'd be no reason to use it in the first place. Let's assume that was the case, then probably 90 to 95 percent of all users in the world could use isolated recognition, economically, vs connected.

Dr. Wayne Lea: Related to that question, at NASA we funded some research in the air traffic control area to investigate how well a person would do in actually performing a task, given that they were doing isolated word recognition, or recognition of speech in restricted format sentences and various other language types. I think that kind of exploratory research can help answer this question. I don't think it's a black and white one, as I think we've seen it alluded to already that in some circumstances you can definitely train users to handle the isolated words. But would they be performing more effectively if they had more? That seems to be a relevant part of the question, and the experiments that were done at Drexel Institute of Technology in the 1960's suggested that when they are given at least a reasonable amount of continuous speech ability then they will do better. So I think that this kind of experiment, if it had been continued more, might have answered this question. Of course, work in the field would answer it, too, if we had the choice, but we don't really have the choice of anything but the isolated recognizers out there right now. I think that even when an isolated word recognizer would have been adequate, one that provided a little more continuous speech would have been better. The ultimate question in many circumstances is, "How is the man-machine combination performing in cost effectiveness?" And I think it goes up with continuous speech in some circumstances.

Mr. Steve Harris: I'm Steve Harris from NAMRL, Pensacola. I'd like to say, first of all, that we have an isolated speech recognition system at NAMRL, and we're quite happy with it because it performed as advertised. One of our applications has to do with research on human information processing limitations and capabilities, and we're interested in just how much control humans have over their own performance. I'm going to make this brief. We've been addressing some of these generic questions, and we've discovered something that's probably intuitively well-known, but at least we've documented it with data. This is, that under some conditions it's very difficult to slow down performance in certain kinds of tasks. If you wish another way to say that, it is hard to be bad in your performance, because there are certain kinds of tasks in which, if you ask subjects to slow down their performance, their performance gets worse. Naturally, they will slow down because that's what you asked them to do, but it also becomes a more difficult task.

There must be a way of categorizing tasks to take advantage of that kind of information, but I think it's a very dangerous sort of thing to make a general statement that isolated word recognition systems would be adequate in 95 percent of the cases. I don't think we have enough information about the limitations of the human's ability to handle many kinds of man-machine systems to make that statement.

Mr. Thomas B. Martin: I would like to say that I think, in the long run, there is no doubt that people will be connected speech oriented, because this is a field where the machine is adapting to the person and not vice versa. You're now trying to remove as many constraints as you can because it's the most natural form of communication there is, that is, speech. At the present time, the matter of economics should determine the issue, except in those cases where nothing will suffice because the job can't be done. But in the long run, even if you had some differences in cost (let's say, a decade from now), there are some people who will not use an isolated word system. In particular, let's take professionals like medical people or lawyers, or the like, who wouldn't spend any time training but who will be glad to talk on a phone, and that's about it. There are probably a range of things that will happen, but if the human has his druthers, he'll take connected speech every time, there's no doubt about it. All the answers have to be looked at as, this is the way it is today, and that's the way it will be tomorrow.

Dr. Edward Huff: One of the reasons that motivated our interest in this topic was reference in the early phases of the conference to the possibility of mis-applying speech technology. Also, of course, our interest in this conference is motivated largely by real time command/control applications. So, I ask you as a corollary to the prior question - do you, the experts, see any place, in any of the applications that we've presented in real-time command/control, where we could be making a fundamental error in judgment in proceeding with isolated word systems or should we be waiting a little bit for more continuous systems? For example, there's some fear on our part that if we make an excursion into the use of voice technology, in this case for cockpit applications, and make a misstep, that it could set us and conceivably others back for some period of time. Unfortunately, that's the way the applications area works. It's not particularly forgiving. That's a tough question to ask, I realize. Are there any takers?

Mr. Thomas B. Martin: I think we've just gone through something like that on the ARPA project where, as a matter of fact, I think the project as a whole set back the field of automatic speech recognition. I'll give you an analogy that probably is controversial, but I feel that the goals of the project were such that they tried to leap a tall building at a single bound, rather than trying to walk up a flight of steps. There were tasks that were very practical and very useful that could have been solved on those dowers that were provided. For example, connected digits would have been helpful to a number of military installations, and it's one step beyond what existed at the time of the ARPA project and its lofty goals. I'm of the philosophy that, looking at progress in fields of science, it's very rare that you get an extraordinary breakthrough with massive funding. Sure, you get breakthroughs now and then that are relatively unplanned, and they are probably in the back room of some laboratory. But take a look at the field of computers which is very analogous in software and hardware. The first computers weren't Atlas and Stretch. They were the INIAK of the University of Pennsylvania. They were relatively incompetent boobs compared to what we have today. The field of progress is one historically where you go one step at a time, and for that reason I feel that we have been through a history in speech that did not promote anything positive overall. A lot of little things came out of it, but I think it was very poor in the sense of how it was organized goal wise.

Dr. Donald Walker: I guess as one of the participants in that project I find myself somewhat puzzled by your characterization of it as having set the field back. I feel the program was a success, and that it did achieve the specifications. There are certainly people who have made the kinds of claims that you have, and you are certainly one of the people who have made those kinds of statements more than others, but I think they are in a sense somewhat puzzling. The particular goals of the project were not the goals of automatic speech recognition, and there were many attempts that we made during the project to try to keep clearly in mind that we were addressing very different kinds of questions, of long range questions, and very much more complex kinds of systems. You may argue that, well, the same amount of money could have been spent on more near term, closer to realization, applications. That is a very different kind of characterization than one that says that the project as a whole was not successful. In fact, my own perceptions of this meeting here, and they may be biased because I do come at it from a particular perspective, is that the ARPA program contributed in good measure to the kind of interest that's reflected in this room. It defines -- moreover, it helps establish -- as other very interesting activities presented here, including Threshold Technology's work over the longest period of time, a farther or long term perspective on the field that we certainly did not have when that project started. And so, I guess, I couldn't disagree more.

Dr. Edward Huff: Well, we certainly have a controversy. However, I would like, if possible, to steer away from that controversy, because our interest in real time applications is not going to be benefited by pursuing it further. I think it's a worthy point to bring up, however, and I will not comment...

Dr. Wayne Lea: I wish to disagree...

Dr. Edward Huff: Dr. Lea, do you feel that my comments are inhibiting the conversation?

Dr. Wayne Lea: I think that this is a crucial point, and I think I agree with Tom, believe it or not. I think that the ARPA project had a tragic effect, in one sense. It's not the fault of the ARPA project; it's the fault of government people at higher levels not having the insight to realize that when one large project is going on that it shouldn't disallow good work of more limited form from going on for immediate needs. And that's the tragedy that happened. I think that there wasn't any insight in the government to say, "Hey, as long as you've got an ARPA project, it still shouldn't be impossible to support something over here, and for somebody from another service to support a $50,000 project over there, et cetera." I think the tragedy is when somebody would say, "Hey, I see all this money being spent, I'll just wait until I see what happens with this ARPA project, and then I'll decide whether you, as my government contract monitor under me, will be able to have any more money.". I don't think that should ever have happened, and I think it did happen, and I think that's the tragedy. I think that the ARPA project had some good impact, but I think you've got to also realize that there were 20 years of research in speech recognition, at least, before that time, and that had a lot of impact. It was good work, and it's the most work that you're hearing about, as well as the applications today. And neither of these should be excluded. However, there was a bad issue concerning that first 20 years, too, and it was well represented by the summary in John Pierce's paper in the Journal of the Acoustical Society, in which he said that it's being run by untrustworthy engineers and mad scientists. The reason for that was because there were a number of isolated projects being done by various people and the wheel was being reinvented over and over again. I know, because I was, at NASA, confronted by many people coming in and telling me they were going to solve the world's problems in speech recognition. They had a great recognizer, and they came in with their suitcases filled with recognizers. Time after time I saw these people come, and they had the "latest technique," which was practically identical to the thing the guy down the street had tried five years before and failed. And I asked them what I think we have to ask every time somebody new comes into this field: "Why do you expect to succeed where others have failed?" And that question is rarely answered by people that are in industry. It should be answered. And people in the government should realize that when one project is going on it doesn't disallow the value of other work that is more limited in scope.

Finally, another aspect of this has to be considered. The ARPA project was done frankly to preserve the artificial intelligence community. It was not done to satisfy work in speech recognition, particularly. And maybe we can criticize that. I feel in some sense that was too bad for speech science. Because it caused some cutback in basic speech research, in some cases, out of groups that had been receptive to speech work going on. But it did intend to demonstrate the utility of artificial intelligence ideas to a task that they thought it would be very much applicable to. That was its original goal, and in that sense it was met, but it had the tragedy that Tom spoke about.

Dr. Edward Huff: I would like to change the topic because there are a few items that we would like to get into before this conference is closed. I'd like to switch over to the area of synthesizers, or automatic speech production systems, and put to the floor the question of the relative advantages or disadvantages of them.

Earlier, we had some commentary, in general, about synthesizers versus what I will call "random access voice playback systems" for lack of a better phrase. Apparently, some feel that it makes a great deal of difference as to the quality of the sound, others do not. Is this a legitimate argument? Does it make a difference in terms of how these systems are being applied, or is it just a matter of arbitrary judgment? Is there any hard evidence available?

Mr. Michael Grady: The only hard evidence I know of in a rigorous sense, I guess, was the work that NASA's done by Carol Simpson, where she did a study of pilot intelligibility of synthesized speech. Clay Coler can probably speak of it better than I. Again, my initial reaction to questions like that is that it's very difficult to answer it without talking about the application for which you want to use the speech.

Mr. Steve Moreland: Steve Moreland, Army Aviation Research and Development Command, St. Louis. I think it does make a difference concerning the application. For example, if we try to use a synthesized voice system in an aircraft, and there are ambient noise levels, as we have with Army helicopters, then speech intelligibility goes way down and the point of having the voice warning system in the first place is totally gone. In other words, if you can't understand what's being said, there's no sense of putting it in the aircraft. So I would say that the degree of speech intelligibility in that application would be very critical.

Dr. Edward Huff: Yes, I can comment myself. Reference was made to some of our work done by Carol Simpson. I don't have these data here, so I'm not completely sure, but I recall that there's a difference between the initial quality of the sound vs its ultimate intelligibility. There's no doubt that the current synthesizers sound different. They have an accent of their own. Certainly if one were exposed to that kind of speech for the first time, particularly in noise, I have no doubt that the intelligibility would be degraded. It's fairly clear to us, however, that that's not a long term effect. That is, when one learns the quality of the sound, however unreal it may be, then it seems to come across fine. But there are those who still argue, and I don't know that our research is complete. I certainly agree with your assertion, but I don't have the data and I don't know that anyone else does either, as to the degree to which intelligibility is interferred with, under arbitrary conditions.

Mr. Steve Moreland: Well, I can give you some firsthand evidence concerning synthesized voice systems. We offered two contractors, about four years ago, an opportunity to put their ideas forward, and develop a system. Using phonetically balanced word lists, which are standard empirically derived speech intelligibility measures, we asked them to allow us to measure the speech intelligibility with people. They did this, using two different approaches, and I certainly hope that the speech intelligibility has improved nowadays, because they came up with about 50 percent and 62 percent respectively. Now, that would be unacceptable in our standard communication systems. Incidentally, we were using phonetically balanced word lists that is the American Acoustical Society's accepted approach for assessing speech intelligibility.

One of the criticisms the contractors had of us at the time was, "Well, we agree that's a good test of speech intelligibility, but on the other hand, why don't you try to use pilot jargon, which is what's really spoken in the aircraft cockpit in the first place?" Well, that was a good point because when you think about it in an airline situation, the pilot jargon is repeated over and over and may be quite similar from one time to another. In the case of an Army helicopter flying in a tactical situation, however, that jargon can vary quite a bit, so I would propose that one of the things that's needed is a good measure of speech intelligibility that we can all accept.

Dr. Edward Huff: One further question, just briefly, when was that work done?

Mr. Steve Moreland: Four years ago.

CAPT Barry McFarland: We in the Air Force just recently went through this exercise for adding a digitized voice system to the F-15 aircraft. In that case we found that the pilot acceptance (we have to get it blessed by the pilot before we can put it in the airplane) was much higher for the digitized voice than it was for the synthesized voice, and the cost difference was negligible. We have gone with the digitized voice, therefore, primarily because of user acceptance.

Dr. Mark Medress: I'd like to support what the Air Force representative just described, as we've had a very similar experience ourselves. We built (digitized) voice response systems for the FAA, because they judged synthesis quality to be unacceptable for the user population that they were trying to interface to, which in some cases was the general public, and in other cases was limited to pilot populations. Since we are obviously involved in speech science and pre-recorded voice response systems are the very least complicated way to get speech out of a computer, one of the things that I feel pretty strongly about is that if you have a voice output requirement that can be well described by highly formatted sentences or statements, and if you don't have a lot of variables in those formats, then you can do a very good job with pre-recorded voice response and get very high quality, very natural sounding speech.

But, you run into a lot of trouble when you don't have such highly formatted situations, or when you have to do a lot of stringing together of individual words. Therefore, I think that the argument that Tom Martin made about continuous speech recognition vs isolated word recognition can be turned around to apply here. It seems clear to me that a very natural sounding and intelligible synthesis system is what's needed in the long run and what will be useful in the long run. In the short term, there are some applications that can best be handled by prerecorded voice, and I think that that will have a positive effect on the user population.

Dr. Edward Huff: Well, I'd like to put in one of my own comments here. I would agree with you largely, except for one argument, which is that in some applications, particularly cockpit applications, it has been said that it would be useful for the pilot always to be able to distinguish what's being spoken to him automatically from what's being spoken by a human. So there seem to be some possible benefits to a distinction, as I believe there is some truth to the argument.

Mr. Rex Dixon: I'd like to remind a few of the people here that in 1964 and 1965, a series of experiments were reported involving the testing of synthesis using phonetically and automatically balanced word lists on a terminal analog synthesizer that had been developed at the San Jose Research Laboratory. It was demonstrated at that time that you could obtain, in the same tough format that is used for testing hearing, the W22 word lists, intelligibility figures with a terminal analog synthesizer in the 80's if listeners were not list-familiar, and in the 90's if they were list-familiar. However, when you went from that discreet word system to continuous speech with true synthesis, the name of the game changed drastically, so that the figures that you obtained using intelligibility testing really were not applicable at all, except, perhaps, for some statements which you could make about things like, it produces b's this well.

Mr. Steve Moreland: Yes, I think, when you go to actual sentences intelligibility would improve, but I think the question here is, what is a good measure if you don't use that? At least to my knowledge we don't have good measures that can be used relative, say, to an articulation index. It's very difficult to have something that everyone accepts as a good speech intelligibility measure, and that's needed, I think.

Mr. Rex Dixon: I really intended to say just the opposite. With speech synthesizers, when you go to the complexity of generating continuous speech, the degradation that takes place from, say, 90 percent intelligibility is very large. Now, it's true that when you're using natural speech you do get this appreciation of intelligibility as a function of going to longer utterances. That's true.

Mr. Steve Moreland: When you go to the total sentence, I think it's been understood that a pilot will pick up context. He may not hear "altitude," but if he knows that it's used in a sentence, he will know that was the word. You pick up some words even though you don't hear them clearly. Therefore, in a pseudo sort of way, maybe the intelligibility improves, but for what you're talking about I wouldn't disagree with you.

Mr. Michael Grady: I might point out to people that are particularly interested, that in the paper that we presented, we traced very briefly the historical perspective on why, in a lot of flight training systems that we did, we moved away from analog voice systems into synthesized voice systems. Frankly, one of the biggest reasons was cost. Remember that this was at a time when the notion of digital techniques for storing analog signals wasn't generally available. A lot of these analog systems weren't sharing the digital electronics explosion that was occurring in the late 60's. So, I refer you to the paper if you're interested in that historical trace.

Dr. Edward Huff: Thank you. Let's see, it is now five to four, and we are going to have to end the workshop period. I think we will have to, even though there were a few questions, including a few that were suggested by members of the audience, that we were not able to get to. The discussion has been very interesting, and we plan to go over it in detail. I'm sure we will then appreciate it even more. However, I think now, at this point, we should turn the discussion to the last topic and spend no more than...yes, sir?

Mr. Michael Nye: Could you at least list the topics that we did not have time to discuss?

Dr. Edward Huff: Yes, sir. I will list them in no particular order. It was suggested we talk about the use of syntax in isolated word recognizers. A related question was "Is there a good way to increase performance in range of application of isolated word recognition systems?" Another question, which is very interesting, "What is the weak link in current operational speech recognition technology- cost, substitution rate, isolated word constraint, speaker dependence, etc?" And then finally, "What is the correct way to develop speech technology?" I guess this last one is more of a philosophical question, having to do, I think, with the long term prospects of the field. That is, should we concentrate on basic speech science, or tune existing technology to applications, and so forth. In certain ways I think that we have, if not directly, at least indirectly, gotten at some of these issues. We had a few other minor questions, but I won't go into them.

So, at this point I would like to redirect the discussion to the question of "what next?" Where are we going? I think that probably the government speakers of the second day, and those related, might just comment first. So, I'll turn the discussion over to Commander Curran, because I believe that he has something to say on this topic.

CDR Mike Curran: Three years ago, many of you may remember, at the end of the nameless ARPA symposium that I mentioned in my introductory remarks three days ago, the cry was for some formal continuous vehicle to bring those involved with voice technology together. And three years later we're sitting here. I think it's our concern that it may be three years hence before we get together again.

In order to please everyone, and comply with strong suggestions of the ARPA symposium, at this meeting we brought together DoD, government and industry. Although we cannot be faulted, I don't think that we feel the same responsibility to now bring the whole community that's interested together at more frequent intervals. But I do personally think we have the responsibility of bringing at least government representation together more often.

There are a number of vehicles that can be employed. There are those with only DoD representation, those with only government representation, and those where we can have government representation with invited industry participation. Before I spring a few suggestions on you, I'd like to hear some of the government people here come up with some ideas on how we can continue to exchange information more frequently than every two years.

Dr. Donald Connolly: It's one of the things that I had hoped would come to pass a little over two years ago. (I knew I was getting old, but I didn't think it was that fast.) I had hoped that it would be something that took place within the government perhaps 8 months or a year after the ARPA Symposium. I do think we can and should meet somewhat more frequently than every two years, certainly within the government. All of us within the government are communicating with some part of the academic and industrial community and I don't know exactly how to characterize it; we're all in the same boat. I think we have very similar interests, and certainly no serious economic interest in the matter, and so I think at least we should get together and find out where we're at from time to time.

CDR Mike Curran: Or at least once.

Dr. Donald Connolly: Once is better than none.

Mr. Rex Dixon: Relative to getting together, for those of you who don't know about the conferences, I'd like to invite all of you to attend the International Conference on Acoustic Speech and Signal Processing that is supported by IEEE. It is going to be held in Tulsa, Oklahoma this spring. Those of you who don't know about it, and would like to find out more, you can contact me at my address. I'll be glad to give you the exact dates and the names of who to contact and so forth. The conference covers a broad range of all of the things having to do with signal processing and speech processing, and it is usually considered to be a very good meeting.

Dr. Wayne Lea: Along that same line, although it's probably too late for anybody to make any plans, a week from tomorrow there is a session of the International Phonetic Sciences Congress meeting in Miami after the Acoustical Society meeting. It will be dealing with the question: "Speech Recognition, What is Needed Now?" That session has seven invited papers and an hour and 10 minutes of open discussion about the question of where should this field be going.

CDR Mike Curran: I'm surprised that the government hasn't been jumped on more today. Only once during this symposium were we indicted for not funding and managing a data base. I'm surprised no one jumped on us asking, "When are we going to get a list of your requirements?" That hasn't come across in the meeting. Let me tell you my concern. From the roster of 110 attendees here today, we have 7 government agencies represented. Within those government agencies we have 21 activities represented, and among industry we have over a dozen who I would consider to be serious industry participants. So I think there is a real need for the government representatives to get together and describe in more detail than in the global manner we have, what our specific requirements are, and what specific applications we have in mind. I can only thank industry for being so kind and not taking us to task for the global terms we have used to describe our requirements and our applications. I think some day you may yet ask that question.

Dr. Mark Medress: It seems to me that this meeting has served in some sense as an opportunity for a lot of us to get to know one another. There are some of us who know each other very well from past work, but there are others who I've met here for the first time. I think it would be unrealistic to expect a lot of very detailed interactions to take place, and I think that without this kind of introduction to one another we couldn't look at the next steps. I agree with your suggestion. I think it would be very, very helpful if the government could do a better job of coordinating its activities, and if we could all do a better job of keeping each other and ourselves abreast of new technical developments. I really believe that this is an area where the technology is driving the applications as much as the applications are pulling the technology up. And I think it's something that requires a lot of cooperative effort for us all to succeed.

Dr. Robert Breaux: Well, in defense, the original interest I had in this meeting was based on the fact that there are so many new players within government who are actively involved in what is a potentially high visibility push for speech technology, that it was time for those people to get to know each other. ARPA did some work earlier, and those people got to know each other. But we needed to spread out that understanding, and so we made an attempt during this conference by having two cocktail hours to allow people to get to know each other, to learn what their interests were, and be as open as we could during those informal meetings. That's the kind of stuff that has to be built up again by the new players.

CDR Mike Curran: Dr. Medress mentioned the word "coordination," and that word sounds very non-threatening. There are people within the government, however, to whom even that word is threatening. But since he brought up the word "coordination," it should be known that words like "coordinating", "advising", and "consulting" make many of us in uniform, and out, cringe. Therefore, I'm surprised that the government representatives here have not been more responsive to my statement, because I thought that they would be cringing at the fact that we would want to coordinate or get together. Notice that I didn't use that word -- you did, Mark.

I think that what we see is a need for more frequent exchange of information. If you use the word "coordinate" (to "coordinate" our requirements, to "coordinate" our efforts), it's obvious from the interest of government and industry here, that there's money involved. It is not a limitless bucket, though. It's obvious to me we've identified a number of technology problems. It's obvious, too, that the majority of the funding to resolve these problems will come from the government. So, I think there is a need for the government to continually keep its house in order, and the most innocuous way that we can all get together without stepping on each other's toes is what we're after. I have a friendly spy in the audience who's come up with a brilliant solution to this problem, and I'd like to ask Commander Norm Lane to present a concept which is being used successfully within the government called "TAG" (Technical Advisory Group). Norm, could you just give us a short explanation of what this is?

LCDR Norman Lane: As Mike indicated, the word "coordination", at least in most of the Armed Services, is a headquarters word that means we take it away from somebody and give it to somebody else. No matter how it works out, that's the answer. If you really want to get together with people and talk about what needs to be done, and discuss what others are doing to save you from having to spend money on it, that is, use other people's technology, the first thing you have to do is not scare people to death. You have to make sure that you are believed when you say that all you want to do is get together and exchange information and talk about things.

We've sort of played around with a concept that, for lack of anything better to call it, we call a Technical Advisory Group, TAG for short. We're in the process of finishing up our charter one right now in Human Factors Engineering, called the HFE TAG. We currently have, I guess, 18 government organizations covering most of the services and one NASA group, and we will probably have more. The general idea of this, in this case of Human Engineering, is to have one master TAG which will sit down and kick around issues and figure out where we need things like specific working groups that are relatively unfettered by organizational bounds. The TAG will then allow these people to get together without having to go all the way up through Secretary level in order to get signatures to be able to sit down and meet. This may be an unusual concept for some of you, but it's very, very difficult for a tri-service working group to get a firm charter to actually go to meetings and do useful things. The closest we've come to developing an ability to do things with Secretary level blessing is this TAG concept. This requires a lot of work and a lot of preparation. We have, however, already chartered or affiliated two groups. One is a specialty group in Test and Evaluation Methods. Another is one in that we call a Workload Coordination Committee. One rule that we have for the TAG is to aim very strongly at what might be called a technical management or working level, with no headquarters personnel and no funding sources, that is, basically people who are doing the work. We find that this is working very well. People at that level will talk to each other, they'll tell the truth whenever possible, and in general we've had some very good interchanges. We have not pushed the concept of joint funding across services, a number of these things are occurring quite naturally. So, in general, we're fairly pleased with the progress.

Mike asked if I would just briefly run through this concept. I think he had some thoughts that this type of concept might work for the voice area. It's a little bit more difficult than that, however, because even in Human Engineering, for those of you who aren't in that area, there is a confusing array of different disciplines. But voice seems to come from even more disciplines, and so there are an awful lot of agencies involved, and an awful lot of specialty topics, and it may be quite difficult, in fact, to develop a TAG. It may be much more difficult to do it than any other way because there are groups like the IEEE group, the NATO group and several others, that are really specialty groups within specialty groups. I'm not a signal processor, and I could go to these meetings and get 10 percent out of what's going on. I probably ought to have somebody explaining the material to me because I'm planning eventually to spend some money on some of that stuff. So I need a translator who will come and talk about what they're doing, including what's being funded at a more general level.

CDR Mike Curran: Thank you so much. The reason I invited Norm Lane is because I think he's been quite successful in implementing the concept of the TAG, and getting working groups together which are getting something out of it. I'd like to give you one more perspective before we push this or any other idea. I'd like to go a step higher, at least within DoD. Apparently DoD is aware that we have finally gotten working troops together in several areas. I'd like to ask Commander Paul Chatelier to try to filter through to us how the higher ups in DoD view the working troops getting together. Is this helping us, or not? Are they for it or against it?

CDR Paul Chatelier: Both the DoD and the Navy have expressed enthusiasms for these working groups. As you know, we refer to them as Technology Advisory Groups which have tri-service and NASA membership. These advisory groups are composed of working level professionals. I feel that this particular advisory group on voice technology will be of great benefit in documenting where the technology is and can go, who the government and industrial players are, and what application options such a technology has available. Technology Advisory Groups allow the workers to cut through the bureaucracy and push advancements into the open. I would encourage the voice technology advisory group to hold meetings such as this on at least an annual basis. Perhaps the next one could concentrate on voice technology as it relates to some specific application options, for example, command and control.

CDR Mike Curran: Thank you, Paul. I think we'll leave that with you as a teaser. I can assure the government people here that we won't spring this concept on you, but we're going to get more information on it and get your reaction to it, because we don't want the opportunity that's been created here to pass.

CAPT Barry McFarland: I have one question from where I sit in the engineering development world. Is the TAG, in general, pretty much restricted to laboratory participation, and how do we in the development world identify what the user's requirements are? I'll say, in contrast, that we have a DOD/NASA/DOT simulation technology working group which is not called a TAG. In this case we involve the laboratories, developers, buyers, and the people that use the devices. Not being familiar with the TAG, is it limited to laboratory participation, and if so, I see some problems with that.

CDR Mike Curran: Let me answer quickly, although I know Norm is far more qualified. The answer to the first question is no, it is not limited to laboratory personnel. The answer to the second question is, how do you include the user community? I don't know if anyone knows the answer to that. I was at a tri-service/Army meeting, for example, where they wanted user input. They counted up all the Army commands. We figured we'd have maybe 400 people sitting in at a small technical group meeting to make sure the total user community was represented. I think that in many ways it's a copout for us in uniform or civil service. Instead of us going and working with the user community to identify their requirements to simply invite the total user community to present their needs. I don't know the answer. Anyway, let's close by saying this. I think we all feel the need for a permanent vehicle so we don't wait another two years, and I can assure you, at least as far as the government's concerned, that the three co-chairmen here will be getting together and passing on information about some proposals. You can look forward to that.

# CLOSING REMARKS

CDR MIKE CURRAN

NAVAL AIR DEVELOPMENT CENTER
WARMINSTER, PENNSYLVANIA

I'll try to make this brief. I promised you a quick review to see whether we met our objectives. Actually, I don't think it's worthy of discussion because of the fact that we have had no drop in attendance in three days, the fact that our cocktail hours were well attended and everyone was talking, the fact that all the coffee has been drunk for three days, and the fact that we haven't been at each other's throats. I think we did accomplish our one purpose of getting all of you together to exchange information.

We, the government, I believe have met our goals, and we've identified more of the players that we thought existed. I think that industry should have met its goals, and that you now know who the other side is. So we all should be relatively happy.

By way of thanks, I want to thank NASA Ames, our host. I think they have been excellent. In fact, they couldn't have been better. I want to thank Dr. Huff, personally, who has made this a relatively simple affair. I want to thank Dr. Breaux, our other co-chairman, who's done so much of the up-front work for this symposium. I want to thank our support people, who helped us to overcome unsurmountable difficulties including the photographers, the projectionist, and Hallie Funkhouser. And last but not least, I want to thank Nancy Frazier from Telcom Systems, Inc., who did the coordination work. Finally, I want to thank you all for attending.

(This page intentionally left blank)

## CONCLUSIONS AND RECOMMENDATIONS[1]

DR. ROBERT BREAUX
NAVAL TRAINING EQUIPMENT CENTER
ORLANDO, FLORIDA

What questions should automatic speech recognition research consider next? The answer depends upon whether your orientation is theoretical, man-machine interface, voice data entry, command and control, training, or a combination. Within the government, the emphasis seems to be upon viable applications, because the next few system applications must prove successful or else management interest may be lost.

Therefore, the major efforts now underway are employing low-cost commercially available isolated word recognition (IWR) hardware; further, it is thought that enhanced, complex algorithms can give this real-time hardware all the sophistication required to fill near-term needs. The important current issue, then, is capacity: Can real-world requirements be met with IWR hardware enhanced by adaptive algorithms for pseudo speaker independence, by syntax for pseudo infinite vocabularies, and by mathematical models for pseudo continuous speech? Is the IWR hardware sufficiently robust that software development has top priority?

The degree of success of current applications over the next three to five years undoubtedly will have a significant impact upon the interest within the government to provide serious, continued support to the general man-machine speech communication research effort. Nevertheless, it is the continuing responsibility of government groups to assist advancement in the state of the art by further refining specific near-term application possibilities, as well as clarifying important research issues. Technology advances will define the next level of realistic applications, which, in turn, will generate new research issues.

One procedure for obtaining a coherent, objectives-oriented approach is to insure continued contact between and among the various government agencies and activities. Technical discussion and exchanges of information can provide the forum for continued interest and support. At the very least, it can give management a technical base from which to plan. A major recommendation, then, which follows from this conference/workshop, is that the government groups take serious steps to ensure the establishment of an intra-government technical organization for continued support of speech research and development.

[1] Editorial comments by Dr. Edward Huff are gratefully acknowledged.

(This page intentionally left blank)

## ATTENDANCE

### GOVERNMENT

| Agency | Separate Activities Within the Agency |
|---|---|
| Air Force | Aerospace Medical Research Laboratory, WPAFB, Dayton, Ohio |
| | Air Force Logistics Management Center, Gunter AFB, Alabama |
| | Rome Air Development Center, Griffiss AFB, Rome, New York |
| | Aeronautical Systems Division/ENECH/ENAMB/AER-EX, WPAFB, Dayton, Ohio |
| Army | Army Aeromechanics Laboratory, R&T Labs, Moffett Field, California |
| | U. S. Army Avionics R&D Command, Ft. Monmouth, New Jersey |
| | U. S. Army Electronics Command, Ft. Monmouth, New Jersey |
| DARPA | Information Processing Techniques Office, Arlington, Virginia |
| FAA | National Aviation Facility Experimental Center, Atlantic City, New Jersey |
| NASA | Ames Research Center, Moffett Field, California |
| Navy | David Taylor Naval Ship R&D Center, Bethesda, Maryland |
| | Fleet Numerical Weather Central, Monterey, California |
| | Naval Aerospace Medical Research Laboratory, Pensacola, Florida |
| | Naval Air Development Center, Warminster, Pennsylvania |
| | Naval Air Systems Command, Washington, DC |
| | Naval Ocean Systems Center, San Diego, California |

| Agency | Separate Activities Within the Agency |
|---|---|
| Navy (cont.) | Naval Personnel R&D Center, San Diego, California |
| | Naval Sea Systems Command, Washington, DC |
| | Naval Training Equipment Center, Orlando, Florida |
| | Naval Underwater Systems Center, New London, Connecticut |
| | Office of Naval Research, Arlington, Virginia |
| | Office of Naval Research, Pasadena, California |
| | Chief, Naval Education and Training Liaison Office, Williams AFB, Arizona |
| | Navy Fleet Material Office, Mechanicsburg, Pennsylvania |
| NSA | Ft. George G. Meade, Maryland |

## INDUSTRY

Analytics, Inc., Willow Grove, Pennsylvania

Applimation, Inc., Orlando, Florida

Balz Enterprises, Maitland, Florida

Bell Northern Research, Palo Alto, California

The Boeing Co., Seattle, Washington

Bolt, Beranek, & Newman, Inc., Cambridge, Massachusetts

Bunker Ramo Corporation, Dayton, Ohio

Centigram Corporation, Sunnyvale, California

Dialog Systems, Inc., Belmont, Massachusetts

General Electric, Utica, New York

## INDUSTRY (cont.)

Gould, Inc., Melville, New York

Haskins Laboratories, Inc., New Haven, Connecticut

Heuristics, Inc., Los Altos, California

IBM, Thomas J. Watson Research Center, Yorktown Heights, New York

IBM Tokyo Scientific Center, Yokyo, Japan

Interstate Electronics, Anaheim, California

ITT Defense Communications Division, San Diego, California

Jet Propulsion Lab, Pasadena, California

Lockheed Missiles & Space Co., Sunnyvale, California

Logicon, Inc., San Deigo, California

Marketing Consultants International, Inc., Hagerstown, Maryland

McDonnell Douglas Electronics, St. Charles, Missouri

MIT, Cambridge, Massachusetts

Perception Technology Corporation, Winchester, Massachusetts

Probe Systems, Inc., Sunnyvale, California

SEMCOR, Inc., Moorestown, New Jersey

Signal Technology, Inc., Santa Barbara, California

Speech Communications Research Laboratory, Santa Barbara, California

Sperry Univac Defense Systems, St. Paul, Minnesota

SRI International, Menlo Park, California

System Control, Inc., Palo Alto, California

System Development Corporation, Santa Monica, California

Systems Research Laboratories, Inc., Dayton, Ohio

Telcom Systems, Inc., Arlington, Virginia

## INDUSTRY (cont.)

Teledyne Ryan Aeronautical, San Diego, California

Telesensory Systems, Inc., Palo Alto, California

Texas Instruments, Inc., Dallas, Texas

Threshold Technology, Inc., Delran, New Jersey

Time and Space Processing, Inc., Cupertino, California

## ACADEMIA

Carnegie-Mellon Univesrity, Pittsburgh, Pennsylvania

University of Southern California, Venice, California

# LIST OF ATTENDEES

ALLEN, Jonathan
Professor
Massachusetts Institute of
Technology
Room 36-575
Cambridge, Massachusetts 02139
(617) 253-2509

ALLISON, Jeffrey B.
U.S. Air Force
Air Force Logistics Management
Center
Gunter AFB, Alabama 36114
(205) 279-4581
Autovon 921-4581

ANDREWS, P.J.
Code 03416
Naval Sea Systems Command
Room 880, Crystal Plaza 6
Washington, D.C. 20362

ASHBURN, LCDR James H.
Naval Air Systems Command
Code 4135A
Washington, D.C. 20361

BALCER, Tom
Manufacturing Engineer
Lockheed Missiles & Space Co.
P.O. Box 504, B151 0/8616
Sunnyvale, California 94088
(408) 742-1143

BALZ, G.H.
Balz Enterprizes
2524 Chinook Trail
Maitland, Florida 32751
(305) 644-3296

BEEK, Dr. Bruno
U.S. Air Force
Rome Air Development Center
Griffiss Air Force Base
Rome, New York 13441
(315) 330-3454
Autovon 587-3454/6213

BREAUX, Dr. Robert
Human Factors Lab - Code N-71
Naval Training Equipment Center
Orlando, Florida 32813
(315) 646-5130
Autovon 791-5130

BROMLEY, Dr. Keith
Naval Ocean Systems Center
Code 8111
271 Catalina Blvd.
San Diego, California 92152
(714) 225-6641
Autovon 933-6641

BRUGLER, J.S.
Telesensory Systems, Inc.
3408 Hillview Ave.
P.O. Box 10099
Palo Alto, California 94304
(415) 493-2626

BURG, John P.
President
Time and Space Processing, Inc.
10430 N. Tantau Ave.
Cupertino, California 95014
(408) 996-2200

CARLISLE, James H.
Assistant Professor
University of Southern
California
11 Ave 24
Venice, California 90291
(213) 396-7507

CARLSTROM, LT COL David
U.S. Air Force
Defense Advanced Research Projects
Agency
Information Processing Techniques
Office
1400 Wilson Blvd.
Arlington, Virginia 22209
(202) 694-8096

CHATELIER, CDR Paul R.
Code AIR-340F
Naval Air Systems Command
Washington, D.C. 20360
(202) 692-7419

CLEVELAND, Ralph
Navy Fleet Material Support Office
Defense Activities, Code 942
Mechanicsburg, Pennsylvania 17055
Autovon 430-3744

COHEN, Dr. Danny
University of Southern
California
Information Sciences Institute
4676 Admiralty Way
Marine Del Rey, California 90291
(213) 822-1511

COLER, Clayton R.
Research Scientist
NASA, Ames Research Center
Mail Stop 239-2
Moffett Field, California 94035
(415) 965-5716

CONNOLLY, Donald W.
Research Psychologist
Federal Aviation Administration
National Aviation Facility
Experimental Center
Atlantic City, New Jersey 08405
(609) 641-8200
Autovon 234-1596

COOPER, Franklin S.
Associate Director of Research
Haskins Laboratories, Inc.
270 Crown St.
New Haven, Connecticut 06511
(203) 436-1774

CURRAN, CDR P.M.
Naval Air Development Center
Technology Development Branch
Code 6041
Aircraft and Crew Systems
Technology Directorate
Warminster, Pennsylvania 18974
(215) 441-2561

DEWING, William
Manufacturing Research Engineer,
Senior
Lockheed Missiles & Space Co.
Box 504
0/86-76, B/182
Sunnyvale, California 94086
(408) 742-7844

DILLARD, Homer E.
Section Manager
McDonnell Douglas Electronics
Box 426
St. Charles, Missouri 93044

DIXON, N. Rex
Speech Processing Consultant
IBM, Thomas J. Watson Research
Center
P.O. Box 218
Yorktown Heights, New York 10598
(914) 945-1780

DODDINGTON, Dr. G.R.
Speech Systems Research
Systems & Information Sciences Lab
Texas Instruments, Inc.
P.O. Box 5936
Dallas, Texas 75222

DRAZOVICH, Bob
Computer Scientist
Systems Control, Inc.
1801 Page Mill Road
Palo Alto, California 94304
(415) 494-1165

DUNN, R.S.
NASA, Ames Research Center
Mail Stop 207-5 HQUSARTL
Moffett Field, California 94035

DUVA, James S.
Head, Human Factors Laboratory
Code N-71
Naval Training Equipment Center
Orlando, Florida 32813
(305) 646-5692

EGAN, Larry
Logicon, Inc.
P.O. Box 80158
4010 Sorrento Valley Blvd.
San Diego, California 92138
(714) 455-1330

HERRON, Emmett L.
Human Factors Engineer
Bunker Ramo Corporation
4130 Linden Ave., Suite 302
Dayton, Ohio 45432
(513) 254-2647

ENEA, Horace
President
Heuristics, Inc.
900 North San Antonio Road
Suite C-1
Los Altos, California 94022
(415) 948-2542

FERBER, Leon A.
Vice President
Perception Technology Corporation
95 Cross St.
Winchester, Massachusetts 08190
(617) 729-0110

FEUGE, Dr. Robert
Logicon, Inc.
P.O. Box 80158
4010 Sorrento Valley Blvd.
San Diego, California 92138

FLEMING, Dr. Robert A.
Engineer Psychologist
Naval Ocean Systems Center
281 Catalina Blvd.
San Diego, California 92152
(714) 225-7372
Autovon 933-7372

FUNKHOUSER, Hallie M.
Technical Assistant
NASA, Ames Research Center
Mail Stop 239-3
Moffett Field, California 94035

GEER, Charles W.
Engineer
The Boeing Co.
P.O. Box 2999
M.S. 82-87 ORG 2-3541
Seattle, Washington 98124

GRADY, Michael W.
Logicon, Inc.
P.O. Box 80158
4010 Sorrento Valley Blvd.
San Diego, California 92138
(714) 455-1330

GUPTA, Gokal
Staff Scientist
Bell Northern Research
3174 Porter Dr.
Palo Alto, California 94304
(415) 494-3942

HADDEN, David
Chief, Computer Techniques and Developments Team
Advanced Systems Design and Development Division
Center for Tactical Computer Science
U.S. Army Electronics Command
Ft. Monmouth, New Jersey 07703
(201) 544-2337
Autovon 995-2337

HANSON, CDR D.C.
Director, Electromagnetic Technology
Office of Naval Research
Code 221
800 N. Quincy St.
Arlington, Virginia 22217
(202) 692-4713
Autovon 222-4713

HARRIS, LT Steve
Naval Aerospace Medical Research Lab
Pensacola, Florida 32508
(904) 452-3656
Autovon 922-3656

HARTMAN, Robert S.
VP Electronics
Gould Inc.
Hydrosystems Division
125 Pinelawn Road
Melville, New York 11746
(516) 293-8116

HERSCHER, Marvin B.
Executive Vice President
Threshold Technology, Inc.
1829 Underwood Blvd.
Delran, New Jersey 08075
(609) 829-8900

HICKLIN, Mary
Logicon, Inc.
P.O. Box 80158
4010 Sorrento Valley Blvd.
San Diego, California 92138
(714) 455-1330

HILGENDORF, LT COL Robert L.
U.S. Air Force
Aeronautical Systems Division/AERS
Wright-Patterson AFB
Dayton, Ohio 45433
(513) 255-3766
Autovon 785-3766

HNAT, Kieffer
President
Telcom Systems, Inc.
2300 So. 9th Street
Arlington, Virginia 22204
(703) 979-8300

HUFF, Dr. Edward M.
Deputy Chief
Man-Vehicle Systems Research Center
NASA, Ames Research Center
Moffett Field, California 94035
(415) 965-5734

JERREHIAN, John
Director of Marketing
Centigram
1294 Hammerwood Ave.
Sunnyvale, California 94086
(408) 744-1290

LANE, LCDR Norman E.
Naval Air Development Center
Code 6041
Warminster, Pennsylvania 18974
(215) 441-2561
Autovon 441-2561

LARR, Robert
Code 8143
Naval Air Development Center
Warminster, Pennsylvania 18974
(215) 441-2720
Autovon 441-2720

LAWSON, CDR Robert F., USN (Ret)
Naval Applications Engineer
Office of Naval Research
Scientific Department
1030 E. Green St.
Pasadena, California 91106
(213) 795-5971
Autovon 360-2432

LEA, Dr. Wayne A.
Research Linguist
Speech Communications Research
Laboratory
800A Miramonte Dr.
Santa Barbara, California 93109
(805) 965-3011

LEINER, Dr. Barry M.
Senior Development Engineer
Probe Systems, Inc.
655 North Pastoria Ave.
Sunnyvale, California 94086
(408) 732-6550

LERMAN, Leon
Lockheed Missiles & Space
P.O. Box 504
Dept. 86-10, Bldg. 153
Sunnyvale, California 94080
(408) 742-2539

LEVIN, Eugene
Chief Engineer
System Development Corporation
2500 Colorado Ave.
Santa Monica, California 90406
(213) 829-7511

LEWIS, Warren
Human Engineering Branch
Naval Ocean Systems Center
Code 8231
San Diego, California 92152
(714) 225-7372
Autovon 933-7372

LINDBERG, Arthur W.
Electronics Engineer
U.S. Army Avionics R&D Activity
DAVAA-S
Ft. Monmouth, New Jersey 07703
(201) 544-4271

LINDENBERG, Klaus
Director of Advanced Systems
Seplimation, Inc.
930 Woodcock Road
Orlando, Florida 32803
(305) 896-0730

LOWERRE, Bruce T.
Computer Scientist
Systems Control, Inc.
1801 Page Mill Rd.
Palo Alto, California 94304
(415) 494-1165

McFARLAND, CAPT Barry P.
U.S. Air Force
Aeronautical Systems Division
ENECH
Wright-Patterson AFB
Dayton, Ohio 45433
(513) 255-4109
Autovon 785-4109

McKECHNIE, Don F.
Research Psychologist
Aerospace Medical Research
Laboratory
Human Engineering Division
Wright-Patterson AFB
Dayton, Ohio 45433
(513) 255-4403
Autovon 785-4403

MALAN, Howard
Teledyne Ryan Aeronautical
2701 Harbor Dr.
San Diego, California 92138
(714) 291-7311

MALECKI, Jerry
Office of Naval Research
800 N. Quincy St.
Code 455
Arlington, Virginia 22217

MARINI, CAPT Ronald J.
U.S. Air Force
Aeronautical Systems Division
AER-EX
Wright-Patterson AFB
Dayton, Ohio 45433
(513) 255-3766
Autovon 785-3766

MARKEL, John D.
President
Signal Technology, Inc.
15 W. De La Guerra
Santa Barbara, California 93101
(805) 963-1552

MARTIN, Thomas B.
President
Threshold Technology Inc.
1829 Underwood Blvd.
Delran, New Jersey 08075
(609) 461-9200

MARTINS, John Jr.
Project Engineer
Naval Underwater Systems Center
New London Laboratory MC 315
New London, Connecticut 06320
(203) 442-0771
Autovon 636-2181

MEDRESS, Dr. Mark F.
Manager, Speech Communications
Sperry Univac Defense Systems
Speech Communications Department
Univac Park, P.O. Box 3525
UOP16
St. Paul, Minnesota 55165
(612) 456-2430

MERCER, CAPT William C.
Chief of Naval Education and
Training Liaison Office
Air Force Human Resource Laboratory
Flying Training Division
Williams AFB, Arizona 85224
(602) 988-2611
Autovon 474-6945

MITCHELL, LT Thomas M.
Naval Air Development Center
Code 604
Warminster, Pennsylvania 18974
(215) 441-2889
Autovon 441-2889

MORELAND, Steve
Commander, U.S. Army Aviation
R&D Command
Attn: DRDAV-EQI
P.O. Box 209
St. Louis, Missouri 63166

MURRAY, Don
Telcom Systems, Inc.
320 West Street Rd.
Warminster, Pennsylvania 18974
(215) 672-6250

NYE, J. Michael
President
Marketing Consultants
International, Inc.
100 West Washington St.
Suite 216
Hagerstown, Maryland 21740

NYE, Dr. Patrick W.
Associate Director of Research
Haskins Laboratories, Inc.
270 Crown St.
New Haven, Connecticut 06511
(203) 436-1774

OBERMAYER, Richard W.
Naval Personnel R&D Center
Code 34
San Diego, California 92152
(714) 225-6617
Autovon 933-6617

O'HAGAN, Bob
Staff Scientist
Bell Northern Research
3174 Porter Dr.
Palo Alto, California 94304
(415) 494-3942

OHKOHCHI, M.
IBM, Tokyo Scientific Center
1-11-32
Nagatatho, Chiyoda-Ku
Tokyo 100 Japan

OSBORN, Robert
VP Engineering
Dialog Systems, Inc.
32 Locust St.
Belmont, Massachusetts 02178
(617) 489-2830

OWENS, LT Jerry M.
Chief, Engineering Psychology
Division
Naval Aerospace Medical Research
Lab
Aerospace Psychology Department
Pensacola, Florida 32508
(904) 452-3281
Autovon 922-3656

PAGE, Thomas W.
Director, National Security Agency
9800 Savage Rd.
Attn: R-54 Page
Ft. George G. Meade, Maryland 20755

PAINE, Dr. Garrett
Jet Propulsion Lab
4800 Oak Grove Dr.
Pasadena, California 91103
(213) 354-4047

PAYNE, Roland
Program Manager
Systems Control, Inc.
1801 Page Mill Road
Palo Alto, California 94304
(415) 494-1165

PERLAKI, Kinga M.
NASA, Ames Research Center
Mail Stop 239-2
Moffett Field, California 94035

PFEIFER, Larry L.
Vice President
Signal Technology, Inc.
15 W. De La Guerra
Santa Barbara, California 93101
(805) 963-1552

PLUMMER, John
Centigram Corporation
1294 Hammerman Ave.
Sunnyvale, California 94086
(408) 744-1290

PLUMMER, Robert P.
Assistant Professor
University of Utah
NASA, Ames Research Center
Mail Stop 239-2
Moffett Field, California 94035
(415) 965-5716

POPKY, Arnold
Threshold Technology Inc.
880 Neptune Ct.
San Mateo, California 94404
(415) 345-8615

POOR, Ernest E.
Naval Air Systems Command
Code NAIR 413B, Room 336
Washington, D.C. 20361
(202) 692-2641
Autovon 222-2641

PORTER, J.E.
Logicon, Inc.
P.O. Box 80158
4010 Sorrento Valley Blvd.
San Diego, California 92138

REDDY, Dr. Raj
Professor
Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213
(412) 578-2598

REINS, E.
U.S. Navy
Fleet Numerical Weather Central
Monterey, California 93940
(408) 646-2681
Autovon 878-2681

SATTERFIELD, Ray
Photographer
Naval Air Development Center
Warminster, Pennsylvania 18974
(215) 441-2660

SHOUP, Dr. June E.
Director
Speech Communications Research Lab
800A Miramonte Dr.
Santa Barbara, California 93109
(805) 965-3011

SIMMONS, John C.
Human Factors Engineering Services
Group
Systems Research Labs, Inc.
2800 Indian Ripple Road
Dayton, Ohio 45440
(513) 426-4051

SKRIVER, Christian
Naval Air Development Center
Code 6041
Warminster, Pennsylvania 18974
(215) 441-2561

SORENSON, Lon
Systems Engineer
SEMCOR, Inc.
Strawbridge Lake Office Bldg.
Route 38
Moorestown, New Jersey 08057
(609) 234-6600

STRAATVEIT, Sverre Nils
Electronics Engineer
Naval Underwater Systems Center
Code 315
New London, Connecticut 06320
(203) 442-0771
Autovon 636-2766

STRIEB, Melvin L.
Program Manager, Human Factors
Analytics
2500 Maryland Rd.
Willow Grove, Pennsylvania 19090
(215) 657-4100

THEIS, Timothy
Electrical Engineer
U.S. Air Force
Wright-Patterson AFB
Aeronautical Systems Division
ENAMB, BLD 20 Aero B
Dayton, Ohio 45449
(513) 255-3023

THEISEN, CDR Chuck
Head, Human Factors Engineering
Division
Naval Air Development Center
Code 604
Warminster, Pennsylvania 18974
(215) 441-2691
Autovon 441-2691

VALONE, Robert M.
Acquisition Director
Naval Training Equipment Center
N231
Orlando, Florida 32806
(305) 646-4416
Autovon 791-4416

VIGLIONE, Sam S.
Interstate Electronics
707 E. Vermont Ave.
Anaheim, California 92803
(714) 772-2811

WALKER, Dr. Donald E.
Senior Research Linguist
SRI International
Menlo Park, California 94025
(415) 326-6200

WALLIS, Ben
Computer Analyst
David Taylor Naval Ship R&D Center
Bethesda, Maryland 20084
(202) 227-1533
Autovon 287-1533

WEINTZ, Walter W.
Lockheed Missile & Space Co.
P.O. Box 504, B150 014701
Sunnyvale, California 94088
(408) 742-1490

WHITE, Dr. George
ITT Defense Communications Division
Aerospace Electronics Components
and Energy Group
4250 Pacific Highway, Suite 224
San Diego, California 92110
(714) 226-0806

WHITTED, Harry A.
Electronics Engineering
Naval Ocean Systems Center
271 Catalina Blvd.
San Diego, California 92152
(714) 225-7631
Autovon 933-7631

WHITTON, I. James
Systems Engineer
General Electric - AES
French Road MD288
Utica, New York 13503
(315) 797-1000

WOLF, Jared J.
Senior Scientist
Bolt, Beranek & Newman, Inc.
50 Moulton St.
Cambridge, Massachusetts 02138
(617) 491-1850

WOODRUFF, Kenneth R.
Senior Scientist - Human Factors
Systems Research Laboratories, Inc.
2800 Indian Ripple Road
Dayton, Ohio 45440
(513) 426-4051

WRIGHT, Robert H.
Engineering Research Psychologist
Army Aeromechanics Lab, R&T Labs
MS 239-2, MVSRD, Ames Research Center
Moffett Field, California 94035
(415) 965-5740

YOUNG, Peggy J.
Telcom Systems, Inc.
2300 So. 9th Street
Arlington, Virginia 22204
(703) 979-8300

ZAWACKI, Robin
Senior Systems Engineer
Jet Propulsion Lab - Pasadena
4800 Oak Grove Dr.
Trailer 1201
Pasadena, California 91103
(314) 354-4880

FRAZIER, Nancy L.
Telcom Systems, Inc.
2300 So. 9th Street
Arlington, Virginia 22204
(703) 979-8300