NOTICE WARNING CONCERNING COPYRIGHT RESTRICTIONS:
The copyright law of the United States (title 17, U.S. Code) governs the making of photocopies or other reproductions of copyrighted material. Any copying of this document without permission of its author may be prohibited by law.

SPEECH-UNDERSTANDING SYSTEMS:
Final Report of a Study Group

A. Newell (Chairman)
J. Barnett
J. Forgie
C. Green
D. Klatt
J.C.R. Licklider
J. Munson
R. Reddy
W. Woods

May, 1971

Published for the Study Group by:
Computer Science Department
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Limited number of copies of this report are available from:

Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Copies of this report are available from:

National Technical Information Service
Springfield, Virginia 22151
Price: Full size copy $3.00; microfiche copy $0.95

The research reported in here was sponsored and supported by the Information Processing Techniques Office of the Advanced Research Projects Agency of the Office of the Secretary of Defense under existing contracts with the respective institutions of the committee members.

ABSTRACT

This report provides an evaluation of the state of the art and a program for research towards the development of speech understanding systems. To assess the possibility of such systems four specific tasks were considered and evaluated. Problem areas are identified and discussed leading to the conclusions on the technical aspects of the study. A possible program for research and development is presented.

TABLE OF CONTENTS

PREFACE

1. CONCLUSIONS AND RECOMMENDATIONS 1.1

2. INTRODUCTION 2.1

3. THE USES OF SPEECH INPUT 3.1

4. ORIENTATION OF THE STATE OF THE ART 4.1

5. TASKS FOR STUDY 5.1

5.1 Querying a Data Management System 5.1
5.2 Data Acquisition of Formatted Information 5.3
5.3 Querying the Operational Status of a Computer 5.3
5.4 Consulting on the Operation of a Computer 5.6

6. ANALYSIS OF THE TASKS 6.1

6.1 System Organization 6.1
6.2 Semantic Level 6.1
6.3 Sentence Level 6.3
6.4 Lexical Level 6.4
6.5 Phonemic Level 6.5
6.6 Parametric Level 6.6
6.7 Acoustic Level 6.7
6.8 Conclusion 6.7

7. TECHNICAL PROBLEMS AND PROSPECTS 7.1

7.1 The Continuous Speech Problem 7.1
7.2 The Multiple Speaker Problem 7.1
7.3 The Speaker Dialect Problem 7.1
7.4 The Environmental Noise Problem 7.2
7.5 The Telephone Problem 7.2
7.6 The Tuneability Problem 7.2
7.7 The User Training Problem 7.3
7.8 The Vocabulary Problem 7.3
7.9 The Syntactic Support Problem 7.3
7.10 The Semantic Support Problem 7.4
7.11 The User Model Problem 7.4
7.12 The Interaction Problem 7.4
7.13 The Reliability Problem 7.5
7.14 The Real Time Problem 7.5
7.15 The Processing Power Problem 7.5
7.16 The Memory Problem 7.6
7.17 The Systems Organization Problem 7.6

7.18 The Cost Problem 7.6
7.19 The Completion Date Problem 7.6
7.20 A Target System 7.7

8. WAYS AND MEANS 8.1

8.1 The Plan 8.1
8.2 Specification of Initial Research 8.3
8.3 Cooperative Endeavor, Control and Public Information 8.5
8.4 Requirements for the Contractors Developing the Target System 8.7

APPENDICES

A1. HISTORY AND STAFFING OF THE STUDY GROUP A1.1

A2. INTRODUCTION TO SPEECH A2.1

A3. DATA ON HUMAN PROCESSING RATES A3.1

A4. VOICE-DM A4.1

A4.1 DS/2 Extensions for Speech Input A4.1
A4.2 Protocol of Voice-DM A4.2
A4.3 Syntax for (Written) DS/2 . . . . . A4.3

A5. VOICE-KP - DESCRIPTION AND EXPERIMENT A5.1

A5.1 System Design A5.1
A5.2 Results of the Experiment A5.5
A5.3 Discussion A5.5

A6. VOICE-CS A6.1

A6.1 The Data Base A6.1
A6.2 Input Vocabulary A6.1
A6.3 Input Syntax A6.1
A6.4 Protocol A6.3

A7. VOICE-CC A7.1

A7.1 Recorded Protocol for Voice-CC A7.1

A8. OTHER POSSIBLE TASKS FOR SPEECH-UNDERSTANDING SYSTEMS A8.1

A9. ANALYSIS OF THE TASKS A9.1

A9.1 System Organization A9.1
A9.2 Semantic Level A9.12
A9.2.1 Simulation of Voice-KP A9.12
A9.2.2 The Real-Time Problem A9.13
A9.2.3 The New Word Problem A9.13

A9.2.4 The Synonym Problem A9.14
A9.2.5 The Verification Problem A9.14
A9.2.6 The User Modeling Problem A9.14
A9.2.7 General Semantics A9.15
A9.3 Sentence Level A9.17
A9.3.1 Review of the State of the Art A9.17
A9.3.2 Unsolved Problems at the Sentence Level A9.20
A9.3.3 Constraints Provided by the Sentence Level A9.20
A9.4 Lexical Level A9.21
A9.4.1 The Large Data Base Problem A9.21
A9.4.2 The Effect of Errorful Phoneme Strings A9.25
A9.4.3 Sources of Knowledge A9.25
A9.5 Phonemic Level A9.28
A9.5.1 The Lexical Segmentation Problem A9.28
A9.5.2 Errors in Phonemic Strings and the Multiple Labels Problem A9.31
A9.6 Parametric Level A9.38
A9.6.1 Parametric Variability Resulting from Segmental Context A9.38
A9.6.2 Parametric Variability Due to Syntactic and Semantic Context A9.39
A9.6.3 Parametric Variability Resulting From Speaker Characteristics A9.42
A9.7 The Acoustic Level A9.44
A9.7.1 The Noise Problem A9.44
A9.7.2 The Characteristics of the Transducer A9.44
A9.7.3 Signal Processing Techniques A9.45

A10. A SIMULATION MODEL FOR PROJECTING THE PERFORMANCE OF SPEECH RECOGNITION SYSTEMS A10.1

A10.1 The Model A10.1
A10.2 Validation of the Model A10.4
A10.3 Conclusion A10.7

A11. PHONEMIC ANALYSIS OF A FREE ENGLISH SENTENCE A11.1

A11.1 Glossary of Words Used in the Analysis A11.1

A12. ALTERNATIVE MANAGEMENT SCHEMES A12.1

BIBLIOGRAPHY

LIST OF FIGURES

Figure No.	Title	Page No.
1.1	Specifications for a Speech-Understanding System	1.2
2.1	Considerations for a Speech-Understanding System	2.2
3.1	Positive Features of the Voice Channel	3.2
3.2	Speeds of Various Channels	3.2
4.1	Vicens-Reddy System	4.2
4.2	Specifications of Vicens-Reddy System	4.4
5.1	Description of Voice-DM	5.2
5.2	Protocol of Voice-DM	5.2
5.3	Description of Voice-KP	5.4
5.4	Protocol of Voice-KP	5.4
5.5	Description of Voice-CS	5.5
5.6	Protocol of Voice-CS	5.5
5.7	Description of Voice-CC	5.6
5.8	Protocol of Voice-CC	5.7
8.1	Plan for Development	8.2
A2.1	Phoneme Classification	A2.3
A2.2	Distinctive Features of the Phonemes of English	A2.4
A3.1	Frequency Distribution of Speech Rates	A3.1
A5.1	Experimental Apparatus	A5.2
A5.2	Data for Input by Subject	A5.3
A5.3	Flowchart of Simulation	A5.4
A5.4	Elapsed Time for Input vs. Error Rate Condition	A5.6
A6.1	Information Provided by the Current SYSTAT Program	A6.2
A9.1	Levels, Their Representations and Sources of Knowledge	A9.2
A9.2	Example of a Finite-State Diagram of State of Voice-CS User	A9.3
A9.3	Examples of Elementary Sentence Form: COUNT ATTRIBUTE(OBJECT) = VALUE	A9.4
A9.4	Dialog from an ELIZA-like System Simulating Voice-CS	A9.6
A9.5	Mechanisms for Voice-CS Levels	A9.7
A9.6	Voice-CS Processing of an Utterance	A9.8
A9.7	Request-System Subgrammar with Implications for Elementary Sentence Form	A9.10
A9.8	A Simulated Protocol Illustrating the Restrictions for Voice-CC	A9.11
A9.9	Examples of Augmented Transition Network Syntax Analysis	A9.18
A9.10	Flowchart of Augmented Network Syntax Analysis Program of Woods	A9.19
A9.11	Cumulative Number of Word Types vs. Number of Tokens	A9.23
A9.12	Distribution by Grammatical Class	A9.24
A9.13	Estimates of Similarity Score Distributions for Several Dictionary Sizes	A9.24
A9.14	Curves of Probability of Correct Word Recognition as a Function of Word Length (in Phoneme) for the case of Independent and Identical Probabilities of Correctness for Individual Phonemes in the Word	A9.26
A9.15	Distributions of Similarity Scores And an Estimated Total Probability of Error for the Voice-CS Vocabulary and its Three Subvocabularies	A9.27
A9.16	The Speech Waveform of "How Are You" shows that Word Boundaries in the Form of Diminished Energy (Silence-like Events) Simply do not Exist	A9.29
A9.17	Tree of Segmentations of Continuous Speech	A9.30
A9.18	Distributions for Simulation of Continuous Speech Segmentation and Estimates of the Probability of a False Match	A9.32
A9.19	Technique for Segmentation Simulation	A9.33
A9.20	Phonemic Analysis of Voice-CS Protocol Fragment	A9.35
A9.21	Technique for Matching Phoneme Array for "USERS" Against Voice-CS Vocabulary	A9.36
A9.22	Results of Matching Phoneme Array for "USERS" Against Selected Voice-CS Subvocabulary	A9.37
A9.23	Sound Spectrograms	A9.40
A9.24	Sound Spectrograms	A9.41
A10.1	Phoneme Feature Weights	A10.2
A10.2	Phoneme Similarity Matrix	A10.3
A10.3	Data Used for Calibration of Model	A10.5
A10.4	Distributions of Rank Orders of Errors	A10.5
A10.5	Distribution of Absolute Scores	A10.6
A12.1	Alternative Project Structures	A12.2
A12.2	Phases of 3-Year and 5-Year Projects	A12.5
A12.3	Four Kinds of Project "Management"	A12.6

PREFACE

This report contains the final conclusions, recommendations and analyses of an ad hoc study group set up in the spring of 1970 to consider the feasibility of developing a system that would recognize speech in order to perform some task -- what we came to call a speech-understanding system. The study group was responsive to a request by Dr. Larry Roberts, Director of the Information Processing Technology Branch (IPT) of the Advanced Research Projects Agency of the Department of Defense. It consisted entirely of scientists and engineers already associated in one way or another with existing research contracts of IPT. A full history of the group and its activities is given in Appendix 1. Briefly, it came into existence at a meeting in Pittsburgh on March 31 - April 1, 1970, and held its final meeting in Santa Monica on July 27 - 29, 1970. This minimal interaction has obviously limited the depth to which the group could penetrate in considering the charge. Still, we think we have arrived at an analysis of some value.

The report begins with a summary of the conclusions and the one recommendation -- to wit, that we approve of the conclusions we have arrived at. Then comes the main analysis, followed by a series of appendices that make available additional detail.

The body of the report is written for someone who has some familiarity with speech and with computers (especially software), but who is not an expert in the areas with which the report deals, viz., speech recognition and artificial intelligence. Since we expect (indeed, hope) the report to be of interest to many people who do not fit this image of our "standard reader," we have added Appendix 2, which gives a brief, but self-contained, introduction to the material under discussion.

The study group focussed on technical issues and the problems of what types of research and development activities appeared to be required. For example, its plan is in terms of activities and equipment, not dollars. Likewise, no assumptions were made or discussions conducted about who the organizations and people might be who would work on such a system. In particular, we did not assume that existing IPT contractors would necessarily be the ones involved in such an effort. Our remarks on this score are limited to a few general observations (e.g., that universities are not appropriate places to do development projects).

A number of people outside the committee gave most generously of their time and opinions, often on very short notice. We would especially like to thank Bob Anderson of RAND, Lee Erman, Donald McCracken and Richard Neely of CMU, Gary Goodman of Stanford U., Jamie Corbonell of BBN, Ben Gold and Carma Forgie of Lincoln Lab, Ken Stevens of MIT, and Max Mathews of Bell Telephone Laboratories. We also thank the three organizations, Bolt, Beranek and Newman, Carnegie-Mellon University, and the Systems Development Corporation, who made their facilities available to us for our meetings. None of these people or organizations, of course, are in anyway responsible for the analyses and conclusions reached by the study group.

We wish to thank Mildred Sisko who not only acted as the secretary to the committee but also typed most of the original manuscript. We also wish to thank Roberta Gray, Gertrude Lazier, Dorothy Josephson, and Charlene Novak for their help in the preparation of the manuscript.

1. CONCLUSIONS AND RECOMMENDATIONS

We were charged with determining the feasibility of demonstrating a speech recognition system with useful capabilities and greater power than current isolated word recognition programs (e.g., Vicens, Gold). We posted a set of initial specifications that would clearly be useful. During the study we developed a second set of specifications. These two specifications are described succinctly, side by side, in Figure 1.1. The exact dimensions of our charge, more complete description of these specifications (including the numbering of the attributes) and a full discussion of the issues can be found in the body of the report. We give here our conclusions and recommendations. These, also, are given in brief. More complete statements, with supporting discussion, occur primarily in Sections 7 and 8.

Conclusion 1: Three years is not enough time to achieve a system with the initial specifications.

(1) The difficulties lie both in needed research and in the creation of organizations qualified to conduct such a development.

Conclusion 2: Five years provides a reasonable chance of success for the system with the final specifications.

(1) The system would be a research prototype, though it would be capable of extensive operation for exploration and testing.

(2) The restrictions in the specification serve to assure:

(a) that information from all levels (acoustic, phonetic, lexical, syntactic and semantic) is available to help determine the final semantic interpretation;

(b) that several sources of potential variation are removed by fiat, and do not have to be dealt with.

Conclusion 3: The specifications are not absolute, but represent the best performance that it is prudent to aim for now. In particular:

(1) The known noise characteristics, variability and bandwidth of the current commercial telephone system, coupled with the unknown effects of these on recognition algorithms, makes it imprudent to specify communication by telephone.

(2) The current state of knowledge in how to interface general syntactic and semantic mechanisms to the lower representations (acoustic, phonetic, lexical), makes it imprudent to go beyond simple ad hoc systems to obtain the required syntactic and semantic support.

Both these limitations to the specifications could possibly be removed by research conducted within the time scale of the system development.

Conclusion 4: The major technical requirements beyond the current art are:

(1) The systematization of a substantial existing body of acoustic-phonetic and phonological rules in a form useful for recognition algorithms, and incorporation of such rules to test their effectiveness.

(2) The construction of at least one round of experimental total systems prior to attempting the system with the final specifications.

The body of the report gives a more complete list of the technical requirements, both expansion of (1) and (2), and additional ones of lesser moment.

Conclusion 5: Success requires widespread involvement by several technical communities (principally from within the computer, speech and communication sciences). In particular, especially for the research and early development aspects, effort and attention must be focussed on the ultimate problem of a speech-understanding system through some form of cooperative and evaluative endeavor.

(1) Critical to this development are adequate, public analyses of the structure, performance and task environment of the various experimental total systems that are constructed. This requires high quality public data, intensive instrumentation of systems, detailed descriptions of task environments and construction of performance models.

Conclusion 6: A two stage effort over five years appears to offer the best chances of success:

(1) Immediate initiation of directed research efforts to make available for recognition existing knowledge on acoustic-phonetic and phonological rules.

(2) Immediate initiation of directed research efforts into the syntactic interface to the rest of the recognition system and into the nature of

Initial Specifications		Final Specifications	
The system should:		The system should:	
(1)	accept continuous speech	(1)	accept continuous speech
(2)	from many	(2)	from many
(3)	cooperative speakers,	(3)	cooperative speakers of the general American dialect,
(4)	in a quiet room	(4)	in a quiet room
(5)	over a telephone,	(5)	over a good quality microphone,
(6)	allowing moderate tuning of the system per speaker,	(6)	allowing slight tuning of thesystem per speaker,
(7)	but requiring only natural adaptation by the user,	(7)	but requiring only natural adaptation by the user,
(8)	permitting a vocabulary of 10,000 words,	(8)	permitting a slightly selected vocabulary of 1,000 words,
(9)	but with strong syntactic	(9)	with a highly artificial syntax,
(10)	and semantic support,	(10)	and a task like the data management or computer status tasks (but not the computer consultant task),
(11)		(11)	with a simple psychological model of the user,
(12)		(12)	providing graceful interaction,
(13)	tolerating less than 10% semantic error,	(13)	tolerating less than 10% semantic error,
(14)	in a few times real time,	(14)	in a few times real time,
(15)	on a dedicated system with 10^8 instructions per second	(15)	
(16)*		(16)	
(17)*		(17)	
(18)*		(18)	
(19)	and be demonstrable in 1973 with a moderate chance of success.	(19)	and be demonstrable in 1976 with a moderate chance of success.

* See Figure 2.1 for the description of all the parameters. Parameters with * were not specified.

Figure 1.1. Specifications for a speech-understanding system

telephone communication, both of which are required to move beyond the target specifications.

(3) Early development of groups capable and willing (potentially) to take on a substantial development effort. Each group is to create at least one experimental version of a total system.

(4) Formation of several interrelated efforts to obtain the required focussing of research and critical evaluation:

  (a) A summer institute at the beginning.

  (b) A steering committee formed from all the efforts engaged in the research.

  (c) The use of the ARPA network.

  (d) Generation of high quality data and description of the task environments.

  (e) Adequate instrumentation to measure the performance of the experimental systems.

  (f) Attempts to model the performance of the total systems.

(5) A major decision point at about two years into the program, with explicit criteria for continuation:

  (a) Has the work on acoustic-phonetic and phonological rules tested out?

  (b) Do the potential contractors have the necessary qualifications:

    (i) Have put together a working total system.

    (ii) Have settled on a parametric representation.

    (iii) Have a detailed task description.

    (iv) Have a detailed systems design.

    (v) Have a proposal for hardware, and if the hardware is new, have a plan to obtain adequate software for it.

  (c) Does the research admit upgrading the specifications:

    (i) Use of telephone system?

    (ii) More ambitious syntax and semantics?

(6) Initiation at the two year point of three year development efforts to produce one (or more) versions of the specified system.

Conclusion 7: A program of the sort outlined will accelerate the development of speech-recognition systems significantly over simply continuing with the present level of research. This acceleration derives primarily from:

(1) The mobilization of the technical community now, rather than later, on a set of scientific analyses which seem necessary to an adequate speech-recognition system.

(2) The availability of high grade public data in quantity on the performance of speech-recognition systems and the nature of the task environments in which they must operate.

(3) A multiplier effect, in which researchers not directly involved in an IPT funded effort will find these scientific problems attractive.

(4) The actual push of the development efforts themselves.

Conclusion 8: Though not part of our basic undertaking, which was to study technical feasibility, we believe that the speech-understanding system proposed would represent a significant step toward a capability of potential use* to the military. It would also represent a significant scientific advance, both in computer science and in speech science.

(1) The proposed system appears to be an appropriate direct step toward systems of increased capabilities. Nothing in its specifications, or in the proposed plan, is a dead end, e.g., a demonstration just for demonstration's sake.

Recommendation: On balance we believe that the program outlined has a high enough chance of success, and of payoff, if achieved, so that we can enthusiastically endorse its pursuit.

---

* Usefulness depends in part on cost-effectiveness considerations, and we have made no attempt to analyze future cost-effectiveness.

2. INTRODUCTION

> Automatic speech recognition -- as the human accomplishes it -- will probably be possible only through the proper analysis and application of grammatical, contextual, and semantic constraints. This approach also presumes an acoustic analysis which preserves the same information that the human transducer (i.e., the ear) does. It is clear, too, that for a given accuracy of recognition, a trade can be made between the necessary linguistic constraints, and complexity of the vocabulary, and the number of speakers.
>
> J. L. Flanagan, Speech Analysis Synthesis and Perception, 1965, p. 163.

Is speech input to computer possible? The question is not well posed. It depends on many things. Consider only the list in Figure 2.1. It seems annoyingly long. But each of the concerns is an essentially independent specification that, even with present knowledge, has a strong effect on the feasibility and performance of any proposed speech recognition system. Down towards the low performance end there are combinations that are not only feasible, but beginning to be commercially advertised (e.g., "voice-button" systems). Up towards the high end the responsible posture is that only after other intermediate steps have been accomplished successfully should an estimate be made.

Thus, to address the question of speech input to computers requires the specification of a range of systems. Dr. Roberts, we are sure, did not intend to lay down a precise specification to the study group, when he urged that it be set up. In fact, he avoided writing anything down. Nevertheless, at the Pittsburgh meeting he was induced to state verbally the class of systems he had in mind. Our rendition of his remarks offers an appropriate initial specification, which we presented already in Figure 1.1. The numbers there correspond to each of the questions in Figure 2.1.

The few missing items were specifications Dr. Roberts happened to leave out. Other things were explicitly removed from the charge. We were not to make a cost-benefit analysis, but rather to address ourselves to the technical issues and to the research-managment issues of the means for attaining such a system, if it seemed feasible. We were not to discuss a practical system, but rather a demonstration system. The initial specifications were meant to assure that the demonstration system would indeed be relevant to attainment of practical systems. The question of voice output from the computer was removed from consideration. No one doubted the useful role it could play, or that it had its own share of technical problems. It was assessed as a separate technical problem, except as it might be related to performing recognition.

The study group set its own limits. It agreed to assess the initial specifications. But it also felt it should extend the time frame and consider variations in the specifications. It should consider additional research, if it was directly related to system feasibility. In agreement with the charge, no cost-benefit analysis should be attempted. Likewise, there should be no emphasis on detailed costs, though the general size and form of a development effort should be explored.

Four assertions, to be taken as assumptions, will make evident the type of study undertaken.

(1) A speech recognition system will have to employ information from all levels -- from the acoustic to the semantic -- to effect recognition. This point is certainly common enough in the speech recognition field. Witness our leading quotation by Flanagan. However, almost no work (with one recent exception) has taken such a view as an operational guide, rather than as a promissory note on future research.

(2) There has been a significant amount of work on many aspects of the problem at all levels of the system, though much of this work, especially at the higher levels (i.e., syntactic and semantic) has not been applied to speech recognition.

(3) However, the issue is not one of surveying and pulling together a scattered literature. The work significant to speech recognition is largely visible in the main stream of work in computer science (especially artificial intelligence, computational linguistics and systems programming), linguistics and speech science.

1.	What sort of speech? (The continuous speech problem)	Isolated words? Continuous speech?
2.	How many speakers? (The multiple speaker problem)	One? Small set? Open population?
3.	What sort of speakers? (The dialect problem)	Cooperative? Casual? Playful? Male? Female? Child? All three?
4.	What sort of auditory environment? (The environmental noise problem)	Quiet room? Computer room? Public place?
5.	Over what sort of communication system? (The transducer problem)	High quality microphone? Telephone?
6.	How much training of the system? (The tuneability problem)	Few sentences? Paragraphs? Full vocabulary?
7.	How much training of the users? (The user training problem)	Natural adaptation? Elaborate?
8.	How large and free a vocabulary? (The vocabulary problem)	50? 200? 1,000? 10,000? Preselected? Selective rejection? Free?
9.	What sort of language? (The syntactic support problem)	Fixed phrases? Artificial language? Free English? Adaptable to user?
10.	What task is to be performed? (The semantic support problem)	Fixed response for each total utterance (e.g., table look up)? Highly constrained task (e.g., simple retrieval)? Focussed task domain (e.g., numerical algorithms)? Open semantics (e.g., dictation)?
11.	What is known psychologically about the user? (The user model problem)	Nothing? Interests? Current knowledge? Psychological model for responding?
12.	How sophisticated is the conversational dialogue? (The interaction problem)	Task response only? Ask for repetitions? Explain language? Discuss communication?
13.	What kinds of errors can be tolerated? (Measured, say, in % error in final semantic interpretation) (The reliability problem)	Essentially none (<.1%). Not inconvenience user (<10%). High rates tolerable (>20%).
14.	How soon must the interpretation be available? (The real time problem)	No hurry (non real time). Proportional to utterance (about real time) Equal to utterance with no delay (real-time).
15.	How much processing is available? (Measured, say, in millions of instructions per second of speech)	1 mips? 10 mips? 100 mips? 1000 mips?
16.	How large a memory is available? (Measured, say, in millions of bits accessible many times per second of speech)	1 megabit? 10 megabits? 100 megabits? 1000 megabits?
17.	How sophisticated is the organization? (The systems organization problem)	Simple program? Discrete levels? Multiprocessing? Parallel processing? Unidirectional processing? Feedback? Feed forward? Backtrack? Planning?
18.	What should be the cost? (Measured, say, in dollars per second of speech.) (The cost problem)	.001 $/s? .01 $/s? .10 $/s? 1.00 $/s?
19.	When should the system be operational?	1971? 1973? 1976? 1980?

Figure 2.1 Considerations for a speech-understanding system

(4) The only significant question to be answered is whether, if a total system were put together, there would be enough information in the system as a whole (and mechanisms to use it) to effect acceptable recognition. In short, can the promissory note mentioned above be cashed?

We call the type of system to be investigated a speech-understanding system. The inclusion of understanding is to distinguish the systems somewhat from speech recognition systems. It does not so much indicate enhanced intellectual status, but emphasizes that the system is to perform some task making use of speech. Thus, the errors that count are not errors in speech recognition, but errors in task accomplishment. If the system can guess (infer, deduce, ...) correctly what the user wants, then its inability to determine exactly what the user said should not be held against it -- even as for you and I.

The only way currently to assess the possibility of such systems is to consider and evaluate concrete proposals. Since the task structure is a significant variable in the performance of the system, we selected four concrete, but different, tasks. For each we considered the possibilities for a speech-understanding system. Needless to say, in the short time available we did not carry through detailed analyses. We did endeavor to discuss questions about the state of the art and the possibilities (or lack thereof) of potential solutions with respect to these specific systems, rather than simply in terms of general capabilities.

This leads to the following organization of the report. First, to orient the reader, we present two brief sections: Section 3 discusses the uses of speech understanding systems; in this section, we made no cost-benefit analysis. But for those who have not thought seriously about the possibilities of speech input to computers, some general discussion of the possibilities seems required. Section 4 describes the current state of the art. In Section 5 we introduce the four tasks selected. Section 6 gives the main analysis of speech understanding systems for these tasks. Working from the current art we attempt to specify system structure and to assess how well the parts might work. The upshot of this section is the identification of a series of problems, which Section 7 attempts to discuss systematically, giving our conclusions on the technical aspects of the study. Section 8, the final one, takes up the design of a research and development effort that might succeed. All of the conclusions summarized at the beginning of the report are stated in the last two sections in expanded form. Some of the detail accumulated by the study group is banished to appendices, rather than being included in the sections themselves.

3. THE USES OF SPEECH INPUT

From the viewpoint of a total computer system, speech is simply one input representation among many for obtaining the information required to accomplish a task. The relevant issues concern the ease (or difficulty) with which the user can encode his information into the input message, the rate at which it can be done, and the probabilities of various errors. Relevant also are the processing costs, in equipment and time, of the computer system to decode the input into an internal representation suitable for accomplishing the task, as well as the probabilities of various errors in interpretation.

The remainder of the report addresses itself to the latter issues involving the computer. This section sketches the role speech input occupies in the ensemble of possible input representations.

Modern man comes equipped to generate information in several representations: spoken natural language; written natural language, both script and printing; selections from discrete possibilities in response to instructions, e.g., pushing buttons, turning rotary switches, checking boxes, fingering keyboards, pointing at locations, responding within specified time intervals; and, finally, the construction of crude line drawings. With moderate amounts of training he can utilize artificial encoding systems and artificial languages. With large amounts of training some of these artificial systems can become natural -- e.g., typing, which for the skilled typist dominates handwriting or handprinting in many tasks.

Each of these channels has distinct properties and whether a particular mode is preferred in a communication situation depends on a number of factors. Let us summarize these factors for speech, introducing the necessary distinctions as we go. Figure 3.1 provides a list.

Natural language speech is the preferred channel for all situations where there is spontaneous generation of information. The conduct of the market place, of social gatherings, of courts, legislatures, conferences, etc., all attest to this. So does the construction, everywhere, of human subsystems to take dictation so that the generator of the information can speak and others (less costly) can produce the written documents.

The data rate of speech is substantially faster than writing. Figure 3.2 gives a few typical rates. Perhaps as important as the rate is that in spontaneous communication with speech the human appears not to be speech limited, but rather thought limited, whereas with writing the opposite is true. That is, a person knows what he wants to communicate faster than he can write it, but not faster than he can say it. Even when saying predigested material, our speech apparatus is never used at close to capacity, at least as we currently know how to measure such capacities. (See Appendix 3 for a discussion of these numbers, with sources.)

When the situation is one of transduction rather than creation -- i.e., when the human must take in information in one channel which he then transforms and emits on a second channel -- the advantage of speech is not nearly so clear. In particular, for most continuous skilled operations a continuous channel is much preferred, e.g., driving a car or an airplane or positioning a physical object. Often, in fact, the human does not know how to communicate over a speech channel for such tasks. But even the tasks of typing and keypunching, when accomplished by highly skilled people, appear approximately competitive.

Other advantages of speech are tied to the current state of the wider technology. One of these is the equipment needed at the input end to transduce the signal. The telephone handset is both ubiquitous and inexpensive compared to, say, teletypes or graphic displays. However, the telephone imposes its own limitations in distortions and S/N attenuation of the voice signal.

To communicate over a channel requires occupying that channel with consequent inability to use it otherwise. Speech is thus a useful channel when the hands are occupied or must be free. Similarly, the use of hands for written (or keyed) communication requires the body to be immobile, at least sporadically. If mobility is desired, then voice communication is useful.

The broadcast character of speech may also be of use, although this appears to be a secondary matter, given the capabilities of modern electronic transmission (both by wire and wave) and the limited range of voice. Where broadcasting does seem useful it is related to mobility, e.g., so a man can move around a room. An outstanding example is communication within a small group, that is, a conference. Here anyone can be the speaker and all others must not only receive the information, but be aware that others have also received it. Speaker-listener roles can be switched quite rapidly (in seconds). However, this latter depends not only on spoken speech, but on visual cues as well (as the limited success of conference telephone calls testifies). Although conferencing is conceivable without auditory communication -- e.g., through a mutually viewed common visual display -- it does not seem very attractive.

1. Most effortless encoding of all output channels.
   (Especially if free language permitted.)
2. Higher data rate than other output channels.
3. Preferred channel for spontaneous output.
4. Does not tie up hands, eyes, feet or ears.
5. Can be used while in motion.
6. Can be used easily in parallel with other channels or effectors.
   (Possibly same as point 5, though possibly independent.)
7. Broadcast over short ranges (tens of feet).
8. Inexpensive and readily available terminal equipment.

Figure 3.1: Positive features of the Voice Channel

1.	Reading out loud	~ 4 words/sec	
2.	Speaking (spontaneously)	~ 2.5 words/sec	
3.	Typing (record)	~ 2.5 words/sec	
4.	Typing (skilled)	~ 1 word/sec	(~ 5 strokes/sec.)
5.	Handwriting	~ .4 word/sec	
6.	Hand printing	~ .4 word/sec	
7.	Telephone dialing (note touch-tone)	~ .3 words/sec	(~ 1.5 digits/sec.)
8.	Mark sense cards	~ .1 words/sec	(~ .5 digits/sec.)

Figure 3.2. Speeds of various channels

4. ORIENTATION OF THE STATE OF THE ART

The state of the art in speech recognition has been adequately summarized by Lindgren (1965), Hyde (1968), and more recently by Hill (1971) and Otten (1971). We provide here an overall orientation by describing a single system, that of Vicens and Reddy (Vicens, 1969).*

This system is substantially the most advanced along the dimension of a total system organization that takes into account information from all levels -- a dimension of central interest to the study. With a few judicious side comments it can serve to indicate both current performance and current system structure for the whole field. Actually, Vicens put together several variants of the system for different tasks. For specificity we concentrate on the one that recognizes continuous speech in a highly constrained language to direct a computer controlled arm at the Stanford AI project to pick up blocks, e.g., PICK-UP THE BIG BLOCK AT THE RIGHT SIDE.

The system is organized in a series of levels, each existing internally as a distinct data representation. Figure 4.1 gives an overall block diagram that shows these distinct representations. The input to the system is the speech signal (amplitude versus time) taken in via a high quality microphone. The first stage of processing converts this to a set of speech parameters: 6 measurements, taken each 10 milliseconds. This stage is performed by a hardware circuit, designed for the system (but standard art). These six parameters (amplitudes and zero-crossings for each of three frequency bands) provide a crude extraction of relevant information from the speech signal, in terms of what we understand of the significant speech parameters to be. Quite sophisticated schemes have recently been constructed for extracting parameters (e.g., Schafer and Rabiner, 1970; Atal, 1971).

The second level is labeled a phonetic-parametric representation. The 10 ms minimal segments are aggregated into larger sustained segments, which average 80 ms in length. These sustained-segments are classified into a phonetic alphabet of some fifteen discrete symbols. The coarseness of this classification (relative to the standard phonemic alphabet of about 40 symbols) reflects the unreliability of classification. However, the phonetic label does not carry all the information, since the averaged value of the six parameters and the duration of the sustained segment are carried along as well. Thus this level has both phonemic and parametric characteristics.

Three passes are required actually to determine the sustained-segments, given the stream of minimal segments, and a number of bits of knowledge about the nature of speech are brought to bear here. The first pass combines together segments that are similar in their features; the second pass divides already combined sequences that contain too much variation over all; and the third pass does some more combining, especially of transitional segments that appear to have no independent significance. After these passes the categorization is made on the basis of the average parameters: first into six categories (vowels, fricatives, nasals, consonants, stops, transitions) after which vowels and fricatives are each subcategorized.

The third level is that of the word. There is a dictionary which contains representations of words as sequences of sustained-segments. The average length is about 5 sustained segments (i.e., about 400 ms). In the system under description, the vocabulary consists of only 16 words. In other variants, testing only isolated word recognition, the vocabulary went as high as 560 words. Multiple copies exist for each word, so that the dictionary can contain several times the number of lexical entries (three to fivefold repetition in practice). Redundancy at this level is a major way of dealing with variation by speaker and by occasion. It bypasses the attempt to apply phonological and co-articulation rules to derive a representation that is context independent.

Given a candidate sequence of sustained segments, the word it represents is determined by a matching procedure against sequences in the dictionary. Even with the small vocabulary (16 words), the use of redundant entries expands the actual size enough so that matching all entries is prohibitive. Thus, there is an initial selection of a subpart of the dictionary, and matching takes place only against all items in the subpart. The selection is based on the pattern of 9 vowels and fricatives, which are the most reliable of the sustained segments. The dictionary is indexed by these patterns so that selection is rapid. Of course, the pattern of vowels and fricatives is not entirely reliable. Thus, it is possible to relax the criteria in order to widen the search for successive trials.

The match of the candidate sequence to the sequence of a dictionary entry is not simple.

---

* Those who find this description assumes too much technical background should read Appendix 2, which supplies (albeit briefly) some of that background.

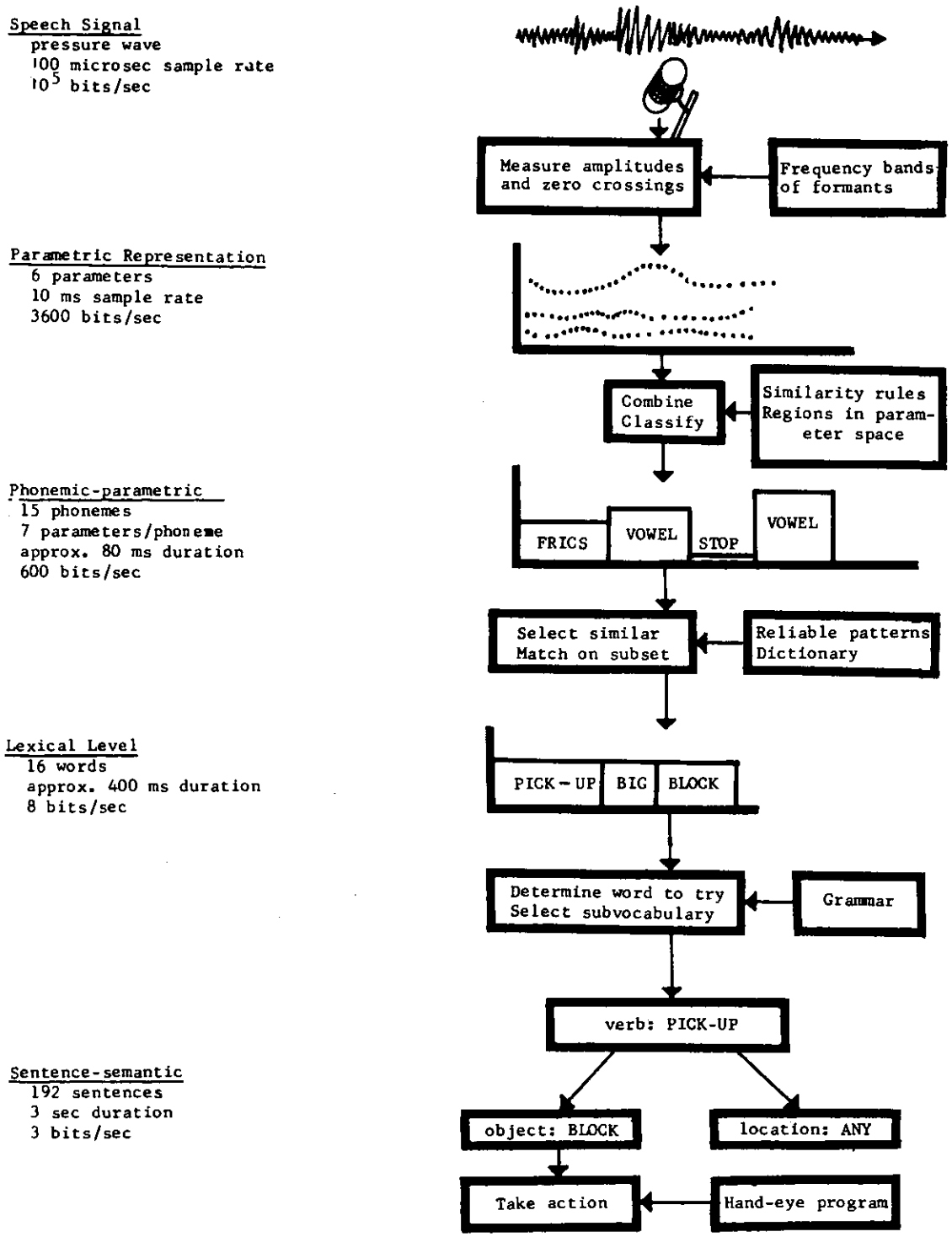


Figure 4.1: Vicens-Reddy System

The sequences need not be in one-to-one correspondence, since various segments could have been missed (or added) in either the dictionary or the candidate sequence (or both). Thus, the matching program, which eventually ends up with a value expressing the degree of match, must consider various possibilities for putting the two sequences in correspondence. This is done again on the basis of the pattern of the vowels and fricatives. The system takes the entry that matches best in the subpart of the dictionary. Absolute levels of confidence are used. The system can decide that the candidate sequence does not correspond to any word at all. This leads to widening the search as mentioned above, or to abandoning the attempt.

The next level is that of the sentence. There is a small grammar of legal sentences and the system assumes that only these are being said to it. It ignores all other words (and sentences). The grammar is highly constrained, admitting only 192 sentences. More important, the grammar exhibits several regularities, which the system exploits to limit its tasks. These regularities are:

(1) A sentence begins either with STOP, RESCAN, or PICK-UP; and only in the latter case does more follow.

(2) If a sentence starts with PICK-UP, then BLOCK occurs somewhere in the sentence.

(3) If a sentence starts with PICK-UP, then the end of the sentence is either the word BLOCK or the words CORNER or SIDE.

(4) Descriptions of the type of block can occur only before BLOCK; descriptions of the location of the blocks can occur only after BLOCK.

We have phrased these rules to make apparent their exploitation in an analysis strategy: Start at the beginning; if the word is not PICK-UP go no further; if it is, find BLOCK wherever it occurs (if BLOCK cannot be found, go pick-up a block at random); search after it for SIDE or CORNER; then find block adjectives (SMALL, MEDIUM, BIG), which precede, and place adjectives (RIGHT, LEFT, TOP, BOTTOM), which follow.

At each stage of the analysis the system goes to the dictionary of words with a different expectation of what subset of words is relevant to the candidate in hand. Thus, at one stage it is considering all possibilities in the phoneme string, but looking only for BLOCK. At other times it is considering several words (SMALL, MEDIUM, BIG) but only searching the phoneme sequence between PICK-UP and BLOCK, both already identified.

The system considers connected speech. However, it does not attempt to detect word boundaries directly. Rather, it identifies syllables (each syllable has one and only one vowel) and matches all possibilities for correspondence between syllables, starting with single syllable words and working up to multiple syllable words. The dictionary, with its representation of each word in terms of a fixed set of syllables, provides a stringent upper limit to the amount of matching required (e.g., BLOCK has only one syllable; BOTTOM may have representations with one or two, but not more, etc.)

There is no semantic level proper in the Vicens-Reddy system, since all the semantic limitations have been built into the structure of the language. The language actually admits a few sentences that the system cannot give semantic meaning to, but when this happens, the system takes an arbitrary action, e.g., the random selection when it can't find BLOCK.

The system accepts multiple users, but requires that they say several passages to the machine in order to obtain the entries for the dictionary. Each word of the legal vocabulary occurs in these passages in a few contexts of legal words and of other words that the speaker might say (e.g., THE, which is not part of the grammar, but often occurs in legal phrases).

With these mechanisms the system can obtain about 85% correct semantic interpretation, corresponding to about 95% correct word recognition. These degrade to about 66% semantic interpretation, corresponding to about 90% correct word recognition, when new speakers use the system calibrated for someone else.

As we noted, Vicens developed variants of the program for other recognition tasks. In isolated word recognition with a single speaker, using the same basic techniques on dictionary selection and matching, the system obtained 90% correct recognition for a single speaker on a 561 word vocabulary. A four-fold repetition of the memory was used. This is the largest vocabulary investigated to date, but other workers have achieved comparable results with vocabularies of 50 to 100 (Bobrow and Klatt, 1968). Vicens also experimented with small numbers of speakers. With a 54 word vocabulary and ten speakers he obtained 80-90% correct isolated word recognition (depending on other conditions). These results are also comparable with other workers (Gold, 1966).

We have not given all the detail of the Vicens-Reddy program, but enough has been exposed to point up several things. First, there are a number of levels in the system, starting with the acoustic and working up to the syntactic and semantics. Second, action is generally from the lower level upward, utilizing programs that incorporate knowledge of that particular level (e.g., that features of speech change slowly on a 10 ms time scale; that vowels are easier to determine than consonants, etc. Third, limitations of the task operate at several levels to help make selections (e.g., the limited subset of words in the dictionary). Fourth, the higher levels are sometimes brought to bear at the lower level in order to make the selection. Thus, the match score of words correctly obtained in the task above cannot be attributed to the lexical level alone, independently of the syntactical level. Similarly, though it does not show in the example above, the phonemic representation is not independent of the lexicon. This feedback can happen when the existence of two closely competing words in the lexicon leads to modification of the phonemic aggregation from the minimal segments, based, say, on additional measurements.

The Vicens-Reddy program hardly exhausts the collection of mechanisms that have been used in speech recognition programs. Every such program has selected for test or exploration only a small set of the mechanisms available. Yet the Vicens-Reddy program can stand as a reasonable statement of the current art, especially given some of the performances from other investigators already quoted. To summarize the program and facilitate evaluation of the specifications given in Figure 1.1, Figure 4.2 recasts the specification of the Vicens-Reddy program in the same forms. Many of the specifications are similiar to those of Figure 1.1 (e.g., item 1). But two or three are severely restricted (e.g., 6, 8 and 9) and these are the price that has been paid for advancing some of the others.

The Vicens-Reddy system:

(1) accepts continuous speech,
(2) from many
(3) cooperative speakers
(4) in a room with 15 db S/N,
(5) over a good quality microphone,
(6) requiring extensive tuning of the system for each speaker,
(7) but no adaptation by the user,
(8) with a carefully selected vocabulary of 16 words,
(9) with extremely strong syntactic semantic support,
(10) and a task of simple commands,
(11) with no model of the user,
(12) or interaction with the user,
(13) at 15% semantic error,
(14) in about 10 times real time,
(15) on a dedicated PDP-10 with 5×10^5 instructions per second,
(16) and 10^6 bits of random access memory,
(17) using a simple program organization,
(18) costing about $3 per second of speech,
(19) and operational in 1968.

Figure 4.2. Specifications of Vicens-Reddy system

5. TASKS FOR STUDY

On initial view the exact task to be performed by a speech-understanding system might not seem critical. In all cases the speech wave must be taken in, and various processes performed on it that seem quite common to all tasks. But such is not the case. A system to recognize the ten digits, spoken in isolation, should capitalize on the fact that it must discriminate one signal from a collection of possibilities, and not from a collection of 150,000 possibilities (all words in a good sized dictionary). To consider questions of vocabulary size, syntactic support and semantic support is precisely to take the exact specifications of these matters as having important effects on the performance of the total system.

What criteria should prospective tasks satisfy? The first 14 dimensions given in Figure 1.1 of the introduction provide one basis (the remainder of the items apply to the system itself, not to the task environment). The tasks should sample variations on these dimensions. We should also consider systems that exploit features that make voice recognition systems interesting from the applied standpoint of communication with computers. Thus, the criteria in Figure 3.1 are relevant.

There is little profit considering tasks that are either well within the current art or far beyond it. For instance a good case can be made for the usefulness of a "voice-button" system, especially in connection with a graphics terminal. The user not only has his hands full, but he wants to communicate commands in coincidence with pointings. To be able to utter one of a small number of isolated commands (copy, erase, print, move, ...) would be extremely useful. But, as we have seen from the brief review of the state of the art, such an application is already technically possible. The issues that still surround it are those of costs versus benefits. Thus, we should not consider this type of system.

On the other hand, a system capable of taking automatic dictation, serving much the same role as a secretary, is too far beyond the current art. Input is in roughly normal speech, with a moderate amount of immediate feedback to correct errors. The system is generally task independent. The actual system might be a "phonetic typewriter," which is able to create essentially a correct transcription, given only the speech signal and phonological knowledge (no higher linguistic or semantic knowledge). Alternatively, it might be a system with extensive knowledge of English language and life. Neither of these seems to us in the realm of possibility, though for distinct reasons. We believe there is not enough information in the speech signal (plus phonology) to permit correct determination of the phonetic transcription. Thus, this system seems not possible in principle.* The other system seems impossible at the moment because the art is not nearly advanced enough. Whether there are essential difficulties, is itself unknown. In any event we do not think it appropriate to discuss technically developing such a system.

We now describe four tasks, starting with the simplest. Figures 5.1 to 5.8 summarize these tasks briefly, listing the features that are of interest to this study. Additional detail can be found in Appendices 4 to 7.

5.1 Querying a Data Management System**

The data management query task (to be done by a system we will call Voice-DM) is to answer questions about files of management information (Figure 5.1). We generated this task by taking an existing system (DS/2, developed at SDC for the IBM 360/30 and /20), and modifying the language just enough to make it verbally conversational. However, we maintained the strict form of the original data language, requiring the speaker to be entirely grammatical within this artificial frame. The query language consists of a finite set of sentence frames which serve to identify a command (print, tally, ...) and to delimit its arguments. The file for the queries is highly organized according to a heirarchy of attributes (equivalently, dimensions or characteristics). The system's response is to be made on an alphanumeric CRT visual display. A fragment of typical interaction is given in Figure 5.2. A more complete description of the DS/2 system and the verbal adaptation is given in Appendix 4. This includes a more extensive protocol, of which Figure 5.2 is simply the first few sentences. This protocol

---

* There is nothing mysterious in this. Humans produce language to be decoded by a system (another human) that takes into account extensive syntactic and semantic constraints. Therefore, humans simply learn to produce speech of a caliber that requires this decoding. Without it there is not enough information in the speech signal.

** An initial description of the task and the protocols in Appendix 4 were provided by Carl Kalinowski of SDC.

The Voice-DM system:

1) accepts continuous speech,
2) from several
3) cooperative, male speakers
4) in a (reasonably quiet) office
5) over a good quality microphone,
7) with natural adaptation by the speaker,
8) with a fixed set of 50 functions words and a large and slowly changing set of content words (~2000)
9) used in a very constrained syntax
10) to do straightforward information retrieval
12) with (CRT) "echoing" by the machine of user input, user editing of the input stream, and optional explanation by the machine of problems encountered,
14) in conversational real-time.

Figure 5.1. Description of Voice-DM

	MACHINE (Display Output)	MAN (Voice Input)
1.		1.
2.	Enter data base information name and volume serial number.	2. PERSONNEL COMMA V50034PERIOD_GO
3.	What is your security key?	3. DEMO_GO
4.	Next;	4. PRINT EMPLOYEES WHERE SEX IS MALE_GO
5.	Undefined print.	5. EQUIVALENCE EMPLOYEES TO EMPLOYEE_GO
	(The plural of employee was not defined in the voice recognition vocabulary; therefore, the PRINT "object" was unrecognized and required definition which was effected by the verbal EQUIVALENCE command.)	
6.	Equivalence employees to employee, Next:	6. PRINT SAME_GO
7.	[Print-out] Next:	

Figure 5.2. Protocol of Voice-DM

(and protocols for the other tasks) were taken by simulating the computer with a human being. The speech was recorded, so that we could examine the type of speech one might expect in the actual situation. Thus, the protocols involve simulated machines, but not canned responses.

The language is highly constrained, having 50 fixed function words (tally, where, go, ...). Within an expression, once one of the words is used (e.g., where), definite expectations exist for which other function words can follow. Thus, there is strong syntactic constraint, which might be convertible to syntactic support for speech recognition. There is also semantic constraint at several levels. A particular installation can be identified by a fixed set of files that it uses, which have a fixed vocabulary of attributes. Though the vocabularies of values (of entries) change slowly over time as new information is added to the file, constancies can occur, e.g., all values of a given attribute are numbers, or Smith, Jones, Thompson, ... are values of the attribute employee indefinitely. During a particular conversation the user enters a context on some particular part of the file and maintains an interest in certain attributes and values over several utterances. Thus, several types of semantic constraint are potentially available to give semantic support to recognition.

The operating environment of Voice-DM is to be similar to current data management systems, i.e., an office. The input to the system can be over high quality microphone and with adequate bandwidth, and the ambient noise level can be kept to tolerable levels. (Having a silent CRT display, as opposed to a printer or typewriter, could be of some help.)

The nature of the task limits the user population to a modest set of moderately trained people, since the data management system itself requires some training and interest. It even seems reasonable to restrict the population to be all of one sex, presumably male. These users could not only be identifiable to the machine, but could themselves adapt somewhat to it, so as to talk in recognizable fashion. This need not be an instructed adaptation, but simply the result of learning during use to avoid errors.

5.2 Data Acquisition of Formatted Information

The second task (Figure 5.3) is the other side of the coin from the one just considered: to enter into a file large amounts of information. We can call the system that does it Voice-KP (for Voice-Key-Punch). The information to be input is in some format, so that only the relevant information, say attributes and values, need be spoken. We can imagine an interaction (either verbal, as in Voice-DM, or written) to set up the formats for communication. Then the human reads in the material, either from a collection of forms or from a table assembled elsewhere. The system presents the input material on a CRT display as it is read in. Written feedback is necessary (not just aural), since whether the spelling is correct is a paramount question. The user monitors the display and corrects any errors. Figure 5.4 gives a fragment of an interaction.

This task seems a plausible candidate to capitalize on the input rates that might be achieved by voice. It is considerably more structured than the full task of taking dictation. However, it still faces some of the same problems, since new material is coming into the system (though some terms will already exist in a dictionary). The attribute-value structure can help, but very little other semantic help exists, since the user is moving from one new item to another. Appendix 5 gives more detail on the task.

5.3 Querying the Operational Status of a Computer

The computer status task (to be done by a system we will call Voice-CS) is to answer questions about the current operational status of a computer system (Fig. 5.5). To be specific we picked the job done in written form by the command SYSTAT on the DEC PDP-10. This prints out a listing of all the resources currently in use or free, the jobs in progress, people on the system, programming systems in use, and various operating statistics. An annotated example of the printout is given in Appendix 6. The user should be able to call Voice-CS on the local telephone and ask it specific questions concerning the status. Voice-CS should understand the question well enough to be able to give the specific answer required (not simply dump the whole data base on the user, as is done by the current SYSTAT). A typical interaction is given in Figure 5.6. Voice-CS must itself communicate verbally. This is a necessary feature of the design, but not one to be considered here. The essential aspect is the user's utterance and the system's decoding from it that information is desired about the loading of the system (as opposed to other items of data in the status log).

The data base from which the system is to work is very small, providing strong semantic restrictions on what can meaningfully be said. On the other hand, the user is to be permitted to utter anything he wants. That is, not only is continuous speech permitted but much broader speech than the system is prepared to handle. Thus, the language specified for the Voice-CS is one that essentially contains only the key words necessary to diagnose the task. In the first interaction in Figure 5.7 loaded and system would be detected and

The Voice-KP system:

1) accepts "pseudo-continuous" speech,
2) from a few
3) cooperative speakers,
4) in a reasonably quiet room
5) over a good quality microphone,
6) requiring extensive tuning of the system for each speaker,
7) and with extensive training of the user,
8) with a large (1,000 - 10,000), mostly free vocabulary,
9) in a rigidly fixed format (which is perhaps also orally specified),
10) and a task requiring mere recording of the input,
12) which requires visual "echoing" and voice correction,
13) <10% error, correctable to ~1%
14) with immediate (~.5 sec) response.

Figure 5.3. Description of Voice-KP

User says:	CRT displays:									
	EMPLNO	SURNAME	INITIAL	SEX	AGE	MARSTAT	DEP	DRAFT	DEG	MAJOR
03284	032A4									
EMPLNO 03284	032A4									
EMPLNO 03284	03T84									
EMPLNO 03284	03T84									
RESET	03T84									
EMPLNO 03284	Z3284									
RESET	Z3284									
EMPLNO 03284	03284									
CALLAHAN	03284	CALLAHAN								
R	03284	CALLAHAN	R							
M	03284	CALLAHAN	R	M						
34	03284	CALLAHAN	R	M	???					
RESET	03284	CALLAHAN	R	M	???					
34	03284	CALLAHAN	R	M	34					
M	03284	CALLAHAN	R	M	34	M				
2	03284	CALLAHAN	R	M	34	M	2			
5A	03284	CALLAHAN	R	M	34	M	2	5A		
MA	03284	CALLAHAN	R	M	34	M	2	5A	M8	
DEG MA	03284	CALLAHAN	R	M	34	M	2	5A	MA	
MUSIC	03284	CALLAHAN	R	M	34	M	2	5A	MA	MUSIC
NEXT										
05289	05289									
COCHRAN-JR	05289	COCHRAN-JR								
C	05289	COCHRAN-JR	C							
M	05289	COCHRAN-JR	C	M						
32	05289	COCHRAN-JR	C	M	32					
S	05289	COCHRAN-JR	C	M	32	S				
0	05289	COCHRAN-JR	C	M	32	S	0			
1Y	05289	COCHRAN-JR	C	M	32	S	0	?????		
DRAFT 1Y	05289	COCHRAN-JR	C	M	32	S	0	1Y		
BA	05289	COCHRAN-JR	C	M	32	S	0	1Y	BA	
MATH	05289	COCHRAN-JR	C	M	32	S	0	1Y	BA	MATH
NEXT										

Figure 5.4. Protocol of Voice-KP

The Voice-CS system:

1) accepts continuous speech
2) from many
3) male speakers
4) in a room with at worst 15 db S/N
5) over local telephone lines using standard telephone handsets
6) requiring some tuning of the system for each user
7) but no adaptation by the user
8) with an unlimited English vocabulary but recognizing only words in a 68-word vocabulary
9) with limited syntactic support
10) and a set of simple questions to be answered
11) with a simple (3-state) model of the user
12) capable of requesting repetitions and requiring affirmations
13) at $< 15\%$ semantic error
14) at 2-5 times real time
15) on a (lightly loaded) time-shared PDP-10 with 5×10^5 instructions per second
16) with 5×10^6 bits of random-access memory.

Figure 5.5. Description of Voice-CS

U: Ah ... ah how many jobs are logged in?
M: 23 jobs are currently logged in.
U: er ah ... I assume the system is up.
M: The system is up.
U: Amazing .. um .. Ok are there any free teletypes?
M: Teletypes 3, 4, 5, 6, 7 and 8 are free.
U: Ok, are there any display lines?
M: Teletypes 3, 4, 5, 6, 7 and 8 are free.
U: Er, Ok, thank you Mr. PDP-10.
M: Are you saying goodbye?
U: Right ...
M: Goodbye.

U - user
M - machine

Figure 5.6. Protocol of Voice-CS

from these two words the request of the user would be inferred. These key words can be defined as a simple language to be used by Voice-CS. Appendix 6 gives the grammar in full. From this grammar one can see that the vocabulary consists of 68 words. In fact, the grammar effectively incorporates all the semantic constraints, since it has been designed to admit those and only those sentences that are semantically meaningful.

Any male speaker who comes on the line is to be accepted so that the population is quite open. Thus, feedback to the user to salvage untoward situations is an eventual aspect of the task. Several pages of protocol are given in the appendix, which make concrete the level of operation that is being specified.

5.4 Consulting on the Operation of a Computer

The last task (done by a system we will call Voice-CC for Voice-computer consultant) is to give specific help to a person attempting to run on a new computer system (Figure 5.7). To be specific we picked the problem of a new monitor system (TENEX) being introduced on the DEC PDP-10 at Bolt, Beranek and Newman. A user, though generally familiar with computers, faces a familiarization task of considerable magnitude in "coming up" on a new system. Rather than do this in the standard (time consuming and frustrating) fashion of reading manuals and trying things out, the user is to operate with Voice-CC as a personal aid. Thus, the user is to be able to frame questions of considerable diversity and sophistication, but strictly about the operational problems of using the PDP-10 via TENEX.

Voice-CC:

1) accepts continuous (1 sentence at a time) speech
2) from many
3) cooperative (male) speakers
4) in a computer (i.e., noisy) room
5) over a good quality microphone,
6) with little training of the system
7) and no training of the user (other than natural adaptation),
8) over a 2 ~ 3,000 word, largely fixed vocabulary,
9) in an artificial, but very English-like language,
10) in order to elicit information in a constrained and fixed domain,
11) for a user with definite questions,
12) in a highly interactive,
13) reasonably correct (< 15% errors)
14) real-time conversation.

Figure 5.7. Description of Voice-CC

Figure 5.8 gives a sample protocol. As with the others, it was done by simulating the machine by a human. In this case the human was one of the programmers responsible for TENEX and the speaker was in fact a user trying to find out how to work within TENEX. The response of Voice-CC is over an auditory channel. The interaction would go too slowly if the human had to stop to read a written channel in the middle of his program specifications. (In all events it would have to be a separate channel from the one that the user is trying to learn how to use.) A more extensive protocol is given in Appendix 7, along with a more complete description of the task.

This task is more ambitious than the other three. It makes use of a sophisticated artificial language that is highly English-like -- essentially at the limits of the current art. The possibility of syntactic support still exists here, since the grammar does in fact constrain what is to happen next. Exploiting it may be a more substantial problem than in the other tasks, with their very elementary languages. Semantic support is also possible, since there is a single focus of the program, involving a fixed set of concepts. (Actually, the concepts are not totally fixed. As any computer user knows, the concepts in an operating system are continually being modified.) However, the semantic structure of the task is more complex than for the other three tasks. On the positive side, there is more local semantic context, in that a user in trying to understand some feature of TENEX remains focussed on a narrow topic for several utterances. In addition, because the user is trying to use TENEX, his direct channel to the PDP-10 provides clues of what he is trying to do and what he might be confused about, hence what might be the content of his utterances.

1. User: How do I log onto TENEX?
2. Computer: Type CONTROL C followed by LOG followed by ALT MODE.
3. User: CONTROL C -- noise -- followed by LOG followed by ALT MODE -- noise -- Now what do I do?
4. Computer: Type your user name followed by ALT MODE.
5. User: Okay.
6. Computer: Beyond this point I will dispense with saying "terminated by ALT MODE." Every executive command in TENEX can be terminated with either ALT MODE or carriage return. ALT MODE will cause printing of any implicit information that belongs on the line while carriage return will not.
7. User: Okay, ah, user name -- noise -- do I have to, ah, put a period or anything?
8. Computer: You can terminate it with a carriage return.
9. User: -- noise -- Okay, what's the password? Ah, carriage return again?

   [The question "what's the password" is directed by the user to himself paraphrasing the current demand by the teletype. A switch could be used to tell the system when it's being addressed.]

10. Computer: Yes.
11. User: -- noise ---------- okay, um, how can I look at my file directory?

. . .

Figure 5.8. Protocol of Voice-CC

General Features of All Tasks

A review of the four tasks will reveal that we did not vary all the task dimensions given in Figure 2.1. In particular all of the tasks insisted on permitting connected speech. One of them (Verbal-KP), it is true, may provide the equivalent of separated speech, since the user naturally tends to pause between entries. But in all cases the user himself is simply asked to behave naturally at the auditory level. However, at the syntactic level we accepted a wide range of artificialities.

No system worked with only a single individual, nor did any require highly trained users. However, some of the tasks consider that the population would be small and sympathetic. One task (Verbal-CS) did open the system up to an indefinite set of users with no good control on the motivation of the users.

On all of the tasks we accepted the occurrence of fragmentary and largely spontaneous speech, though on some (Verbal-DM and Verbal-KP) the structure of the tasks tends to minimize this.

Although one task (Verbal-CS) did use the telephone, the others all were in situations where good quality microphones were acceptable. Noise characteristics of the environment varied, e.g., the Verbal-CC may need to operate in a room with other computing equipment, such as teletypes.

The example tasks do not exhaust the various advantages of voice input, as given in Figure 3.1. We focussed mainly on ease, though in Voice-KP we responded to data rate. In Appendix 8 we list brief descriptions of a few other tasks we considered tentatively before deciding on the presented set.

Finally, these tasks were selected for their analytical value for the study. They are not necessarily tasks of choice for any actual effort, if one were to occur as the result of this study.*

---

* In point of fact, of course, these systems are related to some that are seriously under study by some of us in our capacity as scientists. But this connection reflects only our desire to talk in terms of the concrete situations we most understand, and does not reflect any considered opinion that these systems have any preferred status for other investigators.

6. ANALYSIS OF THE TASKS

We now attempt an analysis of the systems to perform the four tasks just defined. To perform these analyses and comparisons, we have had to make several assumptions about specific approaches to tasks. Other possible approaches to each of these tasks might make the analysis different from what is presented here. Choice of specific approaches to tasks is unavoidable if one wishes to reach the level of detail and concreteness needed to give substance to the general conclusions that are our main objective in this report.

In this chapter we will summarize the results obtained through the analyses and comparisons. More detailed descriptions of the analyses can be found in Appendix 9. In addition, Appendix 10 gives the details of the simulation model that was used in obtaining some of the results.

By analysis we mean a study of the problems that are likely to arise within the confines of the proposed task, the sources of error and the sources of knowledge peculiar to the task, and a few results of simulation studies where these could be obtained without excessive effort. We will raise issues along the way. In the next section we will attempt an orderly summary of these issues, leading to our technical conclusions.

The difficulties in performing such an analysis arise from (1) the total system simply having too many components and interactions for an exhaustive analysis; (2) the requirement, in even a preliminary systems analysis, for more science and engineering than is proper for a feasibility study; and (3) the lack of specific scientific data, which makes certain kinds of estimates impossible. Still, a better feeling for the total problem emerges.

6.1 System Organization

All speech systems constructed to date are organized in levels, corresponding roughly to the levels recognized in linguistic and acoustic research. One such organization appeared in the Vicens-Reddy program. Operationally, a distinct data representation for each level exists in the computer system at run time, and programs operate between levels, taking as input speech represented at one level and producing as output its representation at another level. At each level sources of knowledge about that level of representation must be applied to help determine the actual message being transmitted. We identified several of these sources for the Vicens-Reddy program on the right hand side of Figure 4.1. A speech-understanding system must bring to bear these sources of knowledge on selected samples of data at the various levels of representation. The mechanisms that do this must be embedded in a control structure that sees to the selection of the samples, the activation of appropriate procedures and the storage and retrieval of the various sets of data.

Conceptually, the flow of computation is a linear one through the levels, from the acoustic signal to the semantic level and on to the resulting behavior. The actual flow need not be linear. Feedback from the higher levels to lower ones can lead to a cycle of reprocessing to converge on a recognition. Context-setting from the higher levels to the lower ones can use information from prior processing to select a limited linguistics context within which recognition proceeds. Feed-forward, in which information from lower levels is used to estimate the context at higher levels, can select a restricted context back at the lower level. Error processing can undo prior recognition decisions and take different alternatives. How varied and intensive a collection of such mechanisms can be tolerated depends primarily on the sophistication of the control structure of the system.

Several system organization issues have to be considered for speech-understanding systems. (1) Programs that include all the sources of knowledge will be fairly large. Some form of segmentation and paging scheme will be necessary to make the system work smoothly within the presently available computer systems. (2) If a speech-understanding system takes more than a few seconds to respond to a trivial request, then the user will soon become disinterested in the system. To equal human performance the system must respond to trivial questions as soon as a question is completed. Indeed, sometimes they must be able to answer questions even before they are completed. This implies that the system be sufficiently powerful to perform all the necessary analysis in less than real time. (3) A further implication is that each level does as much analysis as it can as the utterance is being uttered, rather than wait for the completion of the utterance. Subroutines and co-routine mechanisms, the basic organizational structures for computer programs, do not provide adequate control necessary for such a system. It must be possible to interrupt the processes in mid-stream, to preserve their state, and transfer control to other routines at unpreprogrammed points. In effect a fully multi-programmed control structure is required.

6.2 Semantic Level

We begin with the semantic level to consider the problems of representation, the mechanisms used and the sources of error and knowledge. As mentioned earlier, we mainly outline the issues here, leaving more detailed discussion to Appendix 9.

Representation. The following box gives a representation of information at the semantic level for the Voice-CS task. We choose a task to be concrete about the nature of representations.

System's status

- Represented by table in PDP-10, accessible via Monitor
- Fixed structure, known at design time, built into Voice-CS

User's desires for status information

- Represented by an elementary sentence form (hence no separate semantic representation)
- Elementary form fixed by design, built into Voice-CS
- Frequency of requests determined by experience

User's communication state

- Represented by finite state system
- Fixed state system determined by logic of conversation
- Frequency of transitions determined by experience

Similar representations can be formed for the other tasks. All of them will have the following major semantic components: semantics dictated by the task environment, a model of the user, and the dynamic semantics of conversation.

Mechanisms. The mechanisms required for operating on the data at the semantic level will vary from task to task. The following list illustrates the nature of the functions to be carried out.

1. Interpret the sentence form.
2. Make transitions in the user state system.
3. Make transitions in the dynamic semantic system.
4. Permit acquisition of new knowledge.
5. Permit modification of the utterance in the case of errors.

Problems. The semantic level possesses several unsolved problems. Here is a partial list:

(1) The new word problem. Voice-DM and Voice-KP require means for adding new words to the vocabulary. This may be performed by the spelling of these words by voice or through the use of a keyboard (not applicable to Voice-KP).

(2) Model of the user. There have been few attempts at formulating models for predicting user behavior in tasks such as the four we have defined. Work in analyzing information processing of humans (Newell and Simon, 1971; Waterman and Newell, 1971) provides some useful pointers. But there has been essentially no work done on modeling of user behavior for use within speech recognition systems.

(3) The interaction problem. Several different aspects of man-machine interaction must be carefully considered in building speech-understanding systems. Whether a system is useful or not will depend on the grace with which it permits the user to make errors in his conversation. This raises the issue of the use of synonyms in conversation, the problem of verification and correction of a request. The verification and correction may be performed through the use of a visual feedback or a voice feedback like "Did you say..." We have already mentioned the real time interaction problem.

Sources of Error and Knowledge. The Voice-DM, Voice-KP, and Voice-CS tasks perform in restricted environments which make it possible for them to use highly specific semantic constraints in the analysis and recognition of the utterance. For instance, "key word" analysis of the form used in ELIZA-like systems (Weizenbaum, 1966) may be adequate for a system like Voice-CS, and simple fixed formats, for Voice-DM.

Voice-CC, on the other hand, requires the use of powerful semantics for its success. Unlike the other three tasks, it is doubtful whether the Voice-CC can be handled by ad hoc methods of semantic representation. Current work in semantic processing and question-answering systems (Winograd, 1971; Woods, 1970) approaches the generality and complexity required for much of the Voice-CC task, but is still too experimental to consider as useful in the immediate future. These more general semantic systems will of course also be appropriate to the more impoverished situations represented by our other three tasks.

Questions of the type "How do I..." and "What happens if..." are somewhat beyond the bounds of existing question answering systems. It appears that many of these can be conveniently handled in a task-restricted environment such as Voice-CC. However, more research is likely to be necessary in this direction. In the near future it may be necessary to restrict the language of usage, even for Voice-CC. For example, one may have to limit the use of anaphoric references and the use of "this," "that," "it," etc. Another possible restriction might outlaw the use of ill-formed and ungrammatical sentences. If the user asks an ungrammatical question, it may be necessary to ask him to rephrase it.

6.3 Sentence Level

In Voice-DM and Voice-KP the user is normally required to follow a rigid syntax. In Voice-CC and Voice-CS the user can ask questions in a natural language like English. In the case of Voice-CS the fact that the user is dealing with a very limited task domain imposes several contextual constraints (once a key word is recognized within the utterance). In the case of Voice-CC it appears that the spoken sentences are almost always quite simple and short and do not often require the full generality of English grammar.

Representation. The following box illustrates a simple representation that appears adequate for Voice-CS at the sentence level, along with the main sources of knowledge.

Represented by elementary form: (COUNT) ATTRIBUTE(OBJECT)=VALUE
Role of each word (syntax-semantics dictionary) determined by knowledge of English grammar and semantics.
Simple word order rules of English.
Frequency of word orders determined by experience.

All the possible requests for status information can be expressed by filling in (or leaving blank) the four items given by the schema in the table. For instance, MAGTAPE(USER3) = ? requests the magnetic tapes assigned to a specific user, COUNT MAGTAPE(USER3) = ? requests their number, and MAGTAPE(?) = MAGTAPE8 asks who is using a specific magnetic tape. In the case of Voice-DM and Voice-KP the grammars of the language will probably be either finite state of phrase structure grammars and the representation of the sentence level should be similar to that used in programming languages. Representation in the case of Voice-CC probably requires the best that can be done in the way of English grammar, namely some provision for transformation features.

Mechanisms. This level operates on the sentential forms appropriate for each task. Unlike parsing of written text, it may be necessary in the case of spoken utterances to be able to start a parse at any point within the utterance, and parse both backwards and forwards. For instance, if the input to the parser is "? DO ? DO NOW" there is little choice but to attempt analysis of the imperfect "sentence." The following table gives some of the operations that may be performed at the sentence level.

1. Procedures for parsing natural language sentences: e.g., transition network grammars.
2. Procedures for parsing phrase structure grammars.
3. Syntax ambiguity analysis programs.
4. Word boundary ambiguity predictors.
5. Simple word order rules.
6. Parses determined by the user state.

Problems. Two main problems are raised at the sentence level that have to be resolved satisfactorily before we can have sophisticated speech-understanding systems. (1) The problem of parsing in the presence of noise. It seems to be necessary to modify the existing parsers so that they can handle "ah," "er," "um" type interjections and utterances that include false starts. There has been very little work done on this aspect. In addition, any speech-understanding system will introduce noise at each of the levels because of its inability to handle peculiar cases. This might result in insertion or deletion of phonemes from the phoneme string, or incorrect recognition at a lower level. Parsing systems that are capable of handling the above types of noises have not yet been satisfactorily demonstrated, and this is likely to be one of the major bottlenecks in the satisfactory development of working speech-understanding systems. (2) Partial parses. Often people have a tendency to abbreviate sentences leaving out the whole subject or predicate. For example: the partial parses that would result from the protocol in Appendix 6 -- "(Laughter) (b) tough... umm...ok. That is all I want to know... Period. And hanging up will cut off this conversation." It appears necessary that systems should be able to recognize that parsing should be suspended and a new parse attempted of the remaining utterance; then at some point the partial parses should be re-evaluated to see if they should be ignored or combined.

Sources of knowledge. In Voice-CS the user is not required to follow a rigid syntax. However, the system will only answer questions that it understands. The limited task domain imposes several contextual constraints which can be utilized in the analysis of lower levels. For example, given that the key word "job" is recognized, the system should know that it has to be either a question on the status of the job, devices and resources being used by the job, the name of the user running the job, etc. This in turn can permit prediction of the most likely words to be found before and after the key word. The reduction effort resulting from contextual dependency is not known, but it appears to be anywhere from 25% to 75%.

The absence of lexical noise and the use of highly restricted syntax should be of significant help in the case of Voice-DM and Voice-KP. Voice-CC may benefit from the attempt at building natural language parsers. At present we do not know about the power and applicability of these parsing techniques for speech.

6.4 Lexical Level

At this level the size and structure of vocabulary, its internal representation, storage and retrieval of lexical items and the effect of the size of the vocabulary on the response time become important factors.

Representation. The following table provides a representation of speech at the lexical level, along with the main sources of knowledge.

Represented by sequences of words
Finite set of words in dictionary with one (possibly more) phonemic sequence for each (From standard knowledge of English phonetics)
Phonological rules (including conversational transformations)
Stress and intonation rules
Phoneme order statistics A priori from English Calculated for local languages

If the vocabulary can be limited, as it is in our task to various degrees, phonological rules, local stress rules and phoneme order statistics can all become more effective at this level.

Mechanisms. The following table provides a partial list of typical processes that operate on the data at the lexical level.

Preprocessing of input sequence to improve order of alternatives
Using phonological rules
Using language statistics
Sequence of matches of phoneme strings
Search interval in input sequence determined by subgrammar and initial lexical pass
Subdictionary determined by subgrammar
Input phoneme string determined by pauses and reliable phonemes
Subset of subdictionary deter- mined by reliable phonemes of input string
Initial pass to detect clear words, determine estimated user state
Reprocessing of close matches using additional parametric information

Problems. The main problem that arises at the lexical level is the effect of the size of the vocabulary on the time and space required for recognition. In the case of Voice-DM one has to deal with large vocabularies. However, words can be preselected by the designers of the system so as to minimize the search space. In the case of Voice-KP we have to permit greater latitude in what may be uttered and how it is recognized. In the case of Voice-CC one has to permit large vocabularies. How to interpret words that are not known to the system has not yet been explored.

Another important problem at the lexical level is that of evaluating the effect of noise on the recognizability of words. If each phoneme is recognized with the probability p (say, .9), then a word with n phonemes (say, 6) will be correct only with the probability of p^n (i.e., $.9^6 = .53$). The longer the word the greater the likelihood of error. One has to find sources of knowledge at higher levels which can eliminate this exponential degradation in accuracy.

Sources of Knowledge. The ten most used words in the English language account for 50% of all the word tokens that have to be recognized in English language conversation. In the protocols for Voice-CC given in Appendix 7, a total of 430 word tokens were used by the user in ten minutes of conversation; a total of 165 different words were used. If we extrapolate this rate, we would normally expect a working vocabulary of 2,000 to 3,000 words in a Voice-CC-like system.

Given that we have an English language vocabulary of 3,000 words, a useful source of knowledge is the distribution of this vocabulary among various grammatical classes. This would indicate the expected reduction in lexicon search, given that we know the appropriate grammatical class of words to be compared from syntactic consideration. For example, 44% of all the words were nouns, 21% verbs, 22% adjectives, leaving 13% to be distributed among adverbs, pronouns, articles, etc.

A simulation study to consider the effect of increasing vocabulary on the observed phonemic ambiguity was performed. Arguments can be advanced either for the chances of ambiguity falling off or for increasing with increased vocabulary size. In fact, as the vocabulary increases (up to 3,000 words, the limit of the study), the percent of the vocabulary confusable with a given word appears to remain about constant.

An attempt was made to evaluate by means of the simulation model the effect of reducing the Voice-CS vocabulary by the application of semantic criteria. Given that a key word is known, e.g., MAGTAPE, the system can select a subvocabulary concerned with magnetic tapes. The probability of error (confusion of one word with another) was reduced from .31 to .16-.29 in the case of the Voice-CS vocabulary.

An attempt was also made to predict the effect on the amount of search in the dictionary by using phonemic criteria. The dictionary can be partitioned according to the most reliable phonemic characteristics, e.g., the types of vowels and fricatives. When it is known that a word contains a particular pattern of these characteristics, then only the relevant subvocabulary is searched. The simulation studies predicted a reduction of about 60% in the computation effort.

An analysis of the transition network parser was done to estimate the degree of syntactic restriction available to reduce the search at the lexical level. For general English, syntactic constraint seems to be of much more limited value than for restricted languages. The number of words to be considered may be reduced by a factor of 1.5-2 by syntactic subselection in the case of general English, whereas for restricted languages syntactic subselection might reduce the number of candidates by a factor of 10.

6.5 Phonemic Level

At the phonemic level the word is represented by a sequence of phonemes from a given alphabet. Linguists have defined a standard phonemic alphabet which is known to be adequate to represent all words in the English language. However, for recognition purposes, it is not necessary that the phonemes be exactly these standard ones. In most systems it appears that segments with similar features (acoustic or distinctive features) may be used to represent speech at this level without any degradation resulting from this representation.

The following table provides the representation of data at the phonemic level with the main sources of knowledge.

Represented by sequences of phoneme lists, where each phoneme list gives the alternative phonemes that could occur at a given point, ordered in likelihood of occurrence
Parametric representations for each phoneme
Base parametric representations for each phoneme
Co-articulation rules
General rules of continuity for phonemes

Each phoneme might have alternate 2nd and 3rd choices. Also there must be rules for determining the effect of co-articulation within a subsequence of phonemes.

Mechanisms. The mechanisms operating on this level consider various alternate choices of the phonemes to determine the words uttered. Since it is difficult to decide where one word ends and another begins, the problem of lexical segmentation becomes a major issue. The following list gives some lexical mechanisms that operate at this level.

1.	Classification of phonemes
2.	Lexical segmentation procedures
3.	Word boundary ambiguity analyzers
4.	Candidate selection from multiple choice phoneme sequences

Problems. The main unsolved problems at this level appear to be:

(1) Minimization of combinatorial explosion of candidates to be considered because of ambiguity in the labeling of segments.

(2) Lexical segmentation. Here again the problem seems to be one of choosing the right candidates so that one does not have to keep all the possible parses in any given stage of analysis.

(3) Efficient representations for combining common phonemic substrings of different words.

Sources of knowledge. There have been some attempts at lexical segmentation using a dictionary containing phonemic transcriptions of all the words in the vocabulary; that is, match the first part of the string to the dictionary, then segment where the dictionary word says, then match the new first part to the dictionary again, etc. (Reddy and Robinson, 1968). In the case of error-free phonemic sequences, the correct lexical segmentation was determined every time without excessive computation. However, it is not known whether this strategy would work as effectively in the case of errorful strings. The Forgies have attempted multiple labeling of segments using nearest neighbor techniques. A limited simulation study (Appendix 9) shows that almost always the correct choice was made with little combinatorial explosions when the first choice of each segment was used to compare against the whole vocabulary, and then choose only those words with high enough scores for comparison of all other combinations.

Another simulation study (Appendix 9) shows that the expected number of false branches that will survive at each node of the lexical segmentation tree is usually less than one for the tasks considered in this study. The implication is that the combinatorial explosion that might result by keeping multiple choices of words at each node of the lexical segmentation tree may not materialize.

There are also some sources of knowledge for lexical segmentation which result from local clues: (1) certain phonemic sequences cannot occur within a word (Siversten, 1961), (2) suprasegmental features, such as duration, pitch, and amplitude, exhibit different characteristics if there is a word boundary between two segments than if there is not (Lehiste, 1970) and (3) co-articulation effects across word boundaries are much less dominant than within a word (Lehiste, 1964). The main difficulty with these sources of knowledge is that they are in generative form and their analytic counterparts appear to be much harder to formulate.

6.6 Parametric Level

Many different representations of speech at the parametric level have been proposed and tried. Most of these consist of sequences of parallel measurements for each phoneme-sized chunk, be they formants and bandwidths, zero crossings and amplitudes, distinctive features, or ASCON parameters. Other relevant knowledge that has to be represented at this level are rules for the determination of significant parameters of speech, different weights for different features based on the perceptual characteristics of speech. The following table summarizes these:

Represented by sequence of parallel measurements
Articulatory rules for significant parameters of speech
Evidence about perceptual characteristics of speech

Mechanisms. Mechanisms needed at this level include:

(1) Measurement programs for each parameter.

(2) Special measurements under the control of lexical level.

(3) Rules for normalizing for parametric variability resulting from phonetic context, speaker variability, sentential context, etc.

Problems. Major unsolved problems for speech-understanding systems appear at this level. These problems arise when one tries to correct for the effects of phonetic, syntactic, and semantic contexts on the wide variability observed in the parametric representation of the speech utterance. This variability on the parametric level leads to error in all subsequent levels of representation and, in particular, the phonemic level.

Parametric variability resulting from phonetic context is usually explained by considering the complex articulatory gesture that results from the given sequence of phonemes. In general, two articulatory gestures corresponding to two adjacent phonemes is called co-articulation. At any given instant the observed segmental parameters are the direct result of co-articulation of the different gestures.

There have been intensive attempts to predict the effect of co-articulation by means of Acoustic-phonetic rules (Lindblom, 1963; Öhman, 1968; Stevens, House, and Paul, 1966; Broad and Fertig, 1970). These rules are usually in a form suitable for the generation of speech, rather than for analyzing incoming speech. This has led Stevens and Halle (1962, 1964) to suggest "analysis by synthesis" as a model for speech recognition. This model for speech recognition involves a comparison of the input spectrum with some internally generated spectra, and an error signal fed back to the generator for the next stage of analysis-by-synthesis.

If most of these generative rules can also be expressed in an analytic form, then the computationally more economical "hypothesize-and-test" might be more suitable. This technique involves hypothesizing the presence of a phonemic sequence and formulating or selecting a test that would verify the hypothesis. This is one of the methods that has been used successfully in artificial intelligence literature (Newell, 1969).

In the extreme, hypothesis and test may be equivalent to the comparison of spectra in analysis-by-synthesis with no reduction in the computational effort. Usually this is not the case; e.g., it is not necessary to generate the whole formant tragectory when a simpler test of the slope can provide the same information.

Parametric variability resulting from syntactic and semantic context. Segmental parameters are not only affected by phonetic context by also by morphemic, syntactic, and semantic context. Acoustic characteristics of the same word (and thereby the phonemes in the word) can exhibit radical differences depending on the sequence in which it appears. Most of this behavior is rule-governed and, to that extent, can be deciphered from a knowledge of English phonology. As with the phonlogical rules, most rules of this type have been described in articulatory terms and have neither been translated to acoustical implications not tested for their general applicability in the English language.

As mentioned earlier, the problems raised in this section may be critical to successful implementation of any meaningful task dealing with continuous speech. Responsible scientists who are knowledgeable in the area of acoustic phonetics and phonetic structure of English have warned of the extreme complexity of acoustic encodings of phonetic segments. If we do not have satisfactory solutions to the issues raised here, it is doubtful whether we would even have speech-understanding systems which will be applicable in the future to more difficult tasks.

Sources of knowledge. There is a great deal of literature on co-articulation, acoustic-phonetic rules, and phonological rules. However, most of it is in an undigested form with a few aspects that are related to speech recognition embedded within a great deal of information which is not.

6.7 Acoustic Level

The representation at this level may be the original analog signal itself or a sequence of amplitudes of digital wave form. The mechanisms required at this level consist of various signal processing techniques, such as fast Fourier transforms, digital filtering, or analog equivalents of these. These extract the continuous measures which make up the parametric representation.

Problems. The problems at this level deal with techniques of elimination of noise of various types.

(1) Environmental noise. Many speech-understanding systems may have to operate in noisy computer rooms with teletype noises and air conditioning noises distorting the signal. There have been some attempts at noise subtraction (Stockham, 1971) which have worked fairly successfully. The main difficulties appear to be the excessive computation time and the possibility of losing some of the relevant portions of the original signal in the process of noise subtraction.

(2) The other major type of distortion that has to be considered at the acoustic level is that resulting from the use of the telephone as the input device. Unlike noise which adds extraneous information to the signal, the telephone distortion mostly subtracts information at the at the high end and the low ends of the signal. In addition to the bandwidth limitation, there are other types of known signal distortions that take place, such as attenuation distortion, envelope delay, cross modulation, discretization noise, and random noise. There has been no systematic study of the effect of each of these distortions on speech recognition systems. It appears to be imperative that the effect of these distortions be carefully analyzed to study their implications on speech-understanding systems.

6.8 Conclusion

In this section we have followed a path through the various levels of representation that appear to be involved in any speech-understanding system that is to be realized in the next few years. This path has mostly exposed issues and problems, and is a reflection of the more extensive analysis given in Appendices 9 and 10.

For the purposes of the study we need to gather together in an orderly way the various points and opinions that have been raised in the course of the analysis. This is the task of the next section.

7. TECHNICAL PROBLEMS AND PROSPECTS

The preceding section has touched on many technical issues and most of the technical conclusions have already been mentioned in one way or another. However, we need to pull them together to bring out their implications for the ultimate purposes of this study. The basic problems listed in Figure 2.1 provide a suitable framework.

7.1 The Continuous Speech Problem

To summarize: (1) speech recognition must work on continuous speech; (2) the parametric representation of a phonetic element is strongly dependent on the surrounding phonetic context; (3) almost no experience is available with recognition algorithms on connected speech. The variety of distinct significant phonetic contexts (not really estimable, but of the order of thousands) denies any simple data processing or combinatorial solutions (such as keeping data on all possibilities in memory). A few difficult cases might be handled by the phonetic equivalent of an idiom ("how are you" as one word), but this cannot stand as a solution of any generality.

The essential problem of continuous speech is the errors in phoneme-level identification, and not necessarily the problems of segmentation between words. One cannot segment speech into words a priori from an examination of the local character of the signal, except at pauses. However, the essential matching for candidate strings does not have to be done for all phase-shifts of the phonemic string, but only for vowels against vowels. This, coupled with the fact that most words have only one syllable, appears to keep the combinatorics within bounds.

The gap in knowledge about recognition of continuous speech is almost sufficient by itself to force a negative answer to the study until at least one full-fledged intensive effort has been accomplished on continuous speech recognition. The mitigating circumstance is the progress in the last decade in acoustic and phonological theory: (1) lawful relations between the behavior of the articulatory system (as the independent system) and the behavior of the parameters of the speech wave (as the dependent system); and (2) rules for English (or other natural languages) which dictate the phonetic segments that result, given the lexical (phonemic) context. Examples of both types were given in the analysis section.

These laws have not been exploited to any degree in recognition systems. A substantial amount of variation in the parametric representation seems accountable by the proper exploitation of these rules. We have little doubt that this source of information exists. We do have a number of concerns about the completeness of these rules and the degree to which they have been really tested against quantities of speech in varying contexts. Scientific work can be successfully initiated with selected cases and special environments. A speech-understanding system will be forced to take quantities of speech as it comes. Nevertheless, the existing collection of acoustic-phonetic and phonological rules is a major source of knowledge which can be mined in relatively short order (a few years) and moved into a form which can be incorporated into recognition algorithms of various types.

7.2 The Multiple Speaker Problem

It has become traditional to refer to the problem of multiple speakers, partly because a number of earlier speech recognition systems were strongly adapted to a single speaker. Substantial variation occurs in the parametric variation of separate speakers. Existing systems have handled small numbers of speakers (ten), but not with any structural grace. The techniques they used do not permit extension to an indefinite set of speakers.

Some of the variance between speakers can be accounted for by the acoustic-phonetic laws mentioned above. When proper measures are taken for the use of these laws, they become speaker invariant. Additional forms of speaker normalization are undoubtedly possible, and prior success by some of us (Forgie, 1959) and by Gerstman (1968) in finding some normalizations within limited areas when sought gives us some faith in this assertion. Also, careful studies of speaker identification techniques (Wolf, 1970) shed some light on the nature of the problem. Whether the multiple speaker problem goes away as an identifiable problem after such efforts is hard to predict. In any event, special identifiable research does not seem required for this problem, other than the already emphasized work on the acoustic-phonetic rules. The search for normalizations, which is indeed important, seems to us to proceed best in connection with actual (and particularized) recognition systems.

7.3 The Speaker Dialect Problem

This issue contains both questions of dialect spoken and questions of age and sex. Ultimately all questions of speaker variation -- age, sex, person to person, occasion to occasion -- blend into one another. Thus, the exploitation of acoustic-phonetic rules has positive effects on all varieties of this problem. But whereas the issues of multiple speakers must be accepted in any realistic system, the same attitude need not extend to variation of dialect or variation by sex and age. Most of the work on acoustic-phonetic laws and on phonological rules have been taken on what is known as general American dialect. To permit variation in dialects (or accents) would be to make much of this material inapplicable. Similarly, acoustic measurements have often been performed exclusively on adult male speakers, due to the standard analysis tool, the speech spectrograph, being designed to show the format

structure well only in speakers with low fundamental frequencies (i.e., males). This existing data is inapplicable other than to male adults.

Basically, all sources of variation should not be taken on simultaneously. Variations of dialect, age and sex are postponable. They will almost surely yield to developmental effort, if systems become possible for males with the general American dialect. Then, in fact, the motivation will exist to do the additional empirical work to develop corresponding laws and rules for these wider populations.

7.4 The Environmental Noise Problem

Noise of the kinds usually found in natural environments results in performance decrement roughly in correspondence to the amount of noise. There seems to be no reason to put special strictures on environmental noise. When the system's performance is marginal, then the effect of environmental noise, even in modest amounts, will be easily detected. As system margins increase, some of this can be spent, so to speak, in permitting room noise to increase.

The chattering of teletypes or a printer right next to the microphone is another matter, and is not likely to be admissible for some time to come. This may well be true of other types of special noise, though what noise sources are of concern depends on the application.

7.5 The Telephone Problem

Actually, the question is what type of communication channel can be used. However, the possibility of using the ordinary telephone is of overwhelming interest and importance. Its low cost and wide availability are matched by no other input transducer.

The grosser aspects of telephone communication, i.e., the restriction of bandwidth to 300-3000 cycles and the S/N db rating, are perhaps tolerable in the same sense that environmental noise is tolerable. Degradation of performance of a system should be a continuous function of the bandwidth limitation from essentially the full range (50-7000) down to telephone band, and of S/N ratio. However, such gross characterizations are not adequate to describe the effects of communication over the telephone. There are burst noise, distortion, echo, crosstalk, frequency translation, frequency dependent envelope delay, abrupt level change, and clipping of the ends of conversations (on TASI), to list only the aspects currently measured by the telephone system itself.

In addition the carbon button microphone used in current handsets has notoriously unpredictable effects on the speech signal, which may vary from instrument to instrument and from time to time.

We believe it unsafe to rely on using the telephone system. No published results on recognition systems using telephone communication are available at the present. So the position is really one of lack of knowledge. After suitable investigation it could turn out that telephone system noise could join other types of noise, for which there is a fairly dependable degradation of system performace. Thus it could be balanced off against other noise sources and positive margin in designing for a particular application.

The questions surrounding the telephone system seem of finite scope and not conceptually complex. Answers (which could be negative in implication) should be forthcoming for research effort expended. Since the telephone is such an important consideration in the application of voice recognition systems, it seems clear that such research should be done. Even with existing recognition systems, much could be found out at an early date to evaluate the degradation from using telephones.

The common commercial telephone system is not the only communication system of interest. Each such system needs to be investigated in detail. Telephone systems employing pulse-code-modulation (PCM) would appear to have a better chance of being simply a bandwidth limited carrier. Radio communications systems, such as tower-to-plane communications systems, seem to be so noisy and variable by human tests (which form an upper bound for machine performance for some time to come) that they can be dismissed for current work (Beitscher and Webster, 1956).

7.6 The Tuneability Problem

Many current systems respond to the problem of multiple speakers by having extensive information for each speaker, e.g., samples (or multiple samples) of each word for each speaker. This extensive tuning of the system to each user precludes certain kinds of interaction (e.g., that of Voice-CS). Also, as the vocabulary size increases, this kind of strategem must be abandoned in a useful future system.

The burden of an alternative lies with the discovery of normalizations and with the use of acoustic-phonetic rules, already discussed above. If these prove successful, interaction of known character for calibration might still be required to determine speaker dependent parameters. It is not clear why this would be more than a few sentences (depending on the noise being tolerated). Thus, we see this problem as subsidiary to others. However, the variance in the speech signal that can be taken out by the laws and rules currently known is undetermined. Backing up to a more empirical approach may still be required.

7.7 The User Training Problem

The user training problem shows up at several (though no all) levels. The user can modify easily the semantic model that governs his talk. He can modify, though not instantaneously, his lexicon or his syntax. He cannot acquire new phonemes or new phonological rules, at least not without engaging in something as strenuous as learning a foreign tongue without accent. These levels are the means of communication, and such means become increasingly assimilated and automated with time. Taking on new aspects (new words, new grammatical constructions, etc.) is not too difficult, though it requires learning. Modifying already assimilated communication tools is more difficult. When an artificial communication system masquerades as a natural one (i.e., has the form of a natural one, but with added restrictions), the possibilities for successful training are especially suspect (often referred to as the habitability of the language -- Watt, 1968).

This dim view of training the user appears at variance with the common observation of human adaptability -- that the human can learn to work with any machine. However, when a human gets into a situation to which he is not already adapted, he slows down, sometimes considerably. First uses of a programming language are very slow, as is communication with a phrase book in a foreign country. Also, it often takes a long time for a human to adapt, long enough to prohibit such adaptation as a design goal for a system.

The level of adaptation of languages (syntax and vocabulary) should be about the same as that accepted for computer systems. Thus, systems such as Voice-DM are useful designs. It does not seem appropriate to require adaptation of speech production, except what occurs naturally through use of the system. The work of Makhoul (1970) on speaker adaptability to a particular word-recognition system indicates the gains to be made from implicit adaptation could be as much as 5% better sentence recognition.

Our decisions reaffirming only natural adaption by the user in the target system stem primarily from notions of appropriate design. Moreover, we cannot identify any specific problem that training at the speech level would be especially helpful in solving.

7.8 The Vocabulary Problem

Increasing the vocabulary raises the probability of confusions. However, the simple size of a vocabulary is not the most appropriate indicator of its character. Vocabularies are not selected as if they were random samples from fixed universes. As the vocabulary grows, longer words increase. We attempted in the analysis section to estimate the amount of confusability as the vocabulary increases.

Increased confusability is not the only effect of a large vocabulary. The necessity for elaborate matching procedures implies that only a modest number of matches can be made for a given candidate against items in the memory. Exactly how many matches can be afforded depends, of course, on the amount of processing available. But eventually the system must select subsets of the vocabulary for consideration without processing the entire lexicon.

There is no way to know now exactly how these factors will balance out, or which will become critical. A total lexicon of 1000 items seems reasonable. More than this (e.g., the 10,000 suggested in the initial specifications) seems too risky an extrapolation. However, if the whole cluster space grows, there may be no trouble.

How selected the vocabulary can be is important. With large vocabularies substantial numbers of near neighbors and peculiar junctions will occur, so that specially selecting the vocabulary might not seem much use. On the contrary, in any particular application a small number of aspects (words, word-transitions, etc.) will cause a disproprotionate share of the errors. Being able to remove a few specific cases, by a judicious replacement of words, might be worth several percentage points in overall error rates. It seems important to retain mild selective options for the vocabulary, and not strike for completely free (or even perversely selected) vocabularies.

7.9 The Syntactic Support Problem

Throughout we have emphasized the role played by syntactic supports. The issues have been two: how much restriction can be obtained from the syntax; and how that restriction can be utilized. For the simple systems we considered, there is little question on both scores. A few bits of selection per word are available, which could make a substantial difference. And useful techniques exist both at the lexical and at the phonemic level. The simple tasks seem quite within reach from this viewpoint.

The real problem arises in taking the next step to general grammars. This step must be taken, since the range of tasks that admit simple ad hoc grammars is highly circumscribed. For more general grammars there is still considerable uncertainty about how to interface them to the lower speech levels. Our discussion of this in the analysis section (Appendix 9) revealed our current ignorance. However, this ignorance is not due to the existence of hard scientific problems, but to the state of development on computer processable grammars. The field is active and is one of the better understood parts of computer science. But the questions relevant to the intercommunication between syntactic and phonemic levels have not been asked. We suspect that relevant and useful answers would be forthcoming if scientific attention were directed to them.

One should aim for an initial system with a constrained grammar, where the interface can be treated by ad hoc means, since not enough is known to go further. There is a high payoff in getting the requisite research done, and it might even be possible to expand the grammatical component as the result of short run returns from such research.

7.10 The Semantic Support Problem

The example tasks revealed a diversity of semantic aspects. For all but the computer consulting task (Voice-CC) the semantic structure was ad hoc -- whatever structure was required in the computer system to perform the task. In each task we were able to find some aspects of the semantics that were relevant to recognition. In each case we found mechanisms to exploit the semantic knowledge.

The main lesson we learn from these tasks is to formulate the task in a sufficiently explicit and detailed way. The implications for the development of a real system lie not in research on semantics in general, but in great attention to the specifics of the task.

With the computer consultant task the situation changes. The structure we posited for Voice-CC reflects the work on building generalized question answering systems. Here there is a common structure for semantics, and with it the possibility of generalized solution to how the semantic level interfaces with the recognition levels. The model of semantics employed, though still limited in many ways, is entirely adequate to the retrieval and data base tasks typified by the other three examples. Though somewhat overpowering for simple table look-up tasks, such as the computer status task, it provides a way of incorporating models of the user and the conversation that were treated entirely ad hoc in Voice-CS.

Though some mechanisms emerged, such as the dynamic restriction of vocabulary, we were not able to provide a formulation of the semantic interface problem with the same precision that we formulated the syntactic interface problem. The reasons are not hard to find. Though common threads run through the work on semantics, it does not have the structural clarity yet of the work on syntax. The explicit representation for syntax permits the clear statements of general problems that must be faced for all grammars so represented, e.g., the taking of lexical terminals in some order.

We take this failure of formulation to indicate that research on semantics is not yet ready to be locked into work on a speech-understanding system in a direct way. Rather it should be pursued at a more general level until some research occurs (it would only take one good piece) that clarifies how to bring the general formulations to bear.

7.11 The User Model Problem

Even in the simplest tasks we could distinguish clearly a model of the user, with his desires and knowledge. In none of our examples did these models get complex, though they did include explicit rules of social intercourse. Though necessary for handling sophisticated dialogue (see section 7.12), the real reason for including them is the amount of selectivity they provided on the nature of the current utterance.

It is certainly appropriate to call for work to produce better models of the user. Work in psychology has not produced such models, and it cannot be expected to do so without something focussing interest on the problem. However, the history of modeling with respect to man-computer communication generally does not provide much basis for hope. Some feeling for the possibilities is sketched in Appendix 10. The task used there is entirely different from those considered in the study, but one where sufficient is known about human behavior to have a full-fledged model of the user's communicative behavior.

So little has been done with explicit user models that we hesitate to put weight on them. We expect that they will emerge as important for proposed systems, but in an ad hoc form, and will then permit suitable formulation as a deliberate area of research and development for the succeeding generation of speech-understanding systems.

7.12 The Interaction Problem

A key determiner of the total success of the systems we discussed is how skillfully they handle interaction with the user. As we note below (section 7.13), performance is not defined by the correctness of a particular utterance given in response to a request for information, but by whether the user leaves the encounter with the system with the correct information and in a suitable (positive) frame of mind. These characteristics of system performance are not determined by the probabilities of error of semantic interpretation alone.

Successful interaction requires primarily a model of the user that permits the appropriate distinctions. The actions to be taken are not difficult. Unfortunately, the problem is not just one of model building, but also of empirical facts. However, first order models can be built just from our own participatory role in human affairs. We do not think special research can be formulated. However, relatively high aspirations should be held about the kind of graceful interaction the system should perform, though we make no specification of the exact varieties of interaction which should be included. Experience is the best guide to that.

7.13 The Reliability Problem

Our emphasis on the total system implies that the error of primary concern is on the semantic interpretation taken by the computer system in response to an utterance (or utterances) input to it. The level of error that can be tolerated at this final level depends on at least three almost independent things: (1) the task; (2) the psychology of the user; and (3) the alternative input media available.

For a specific task embedded in a specific larger system acceptable error rates can sometimes be estimated. In our tasks, which are taken in isolation, this is not possible. All of our tasks are constructed so that errors of the joint user-computer system are rare. Instead, errors of the computer are apparent to the user and are corrected by additional communication. Thus, the governing factor is the psychological response of the user. If the error rate gets too high, then the user becomes dissatisfied with the response of the system.

Unfortunately, almost nothing is known about user responses to computer systems, either in terms of measures of psychological response or in terms of dysfuntional behavior. This holds for almost all human factors questions with respect to computers. It seems excessive for us to insist on good knowledge of user reactions to speech recognition of various quality. Nevertheless, it would be nice to have. From casual observation (e.g., in the Voice-KP task), errors of 20% seems excessive. It seems plausible to ask for 90% correct semantic response in the sorts of tasks we have considered, but we have little basis for such a figure.

The third factor listed above was the error characteristics of competing communications channels. It is part of the lore about human adaptability that, if necessary, humans will endure incredibly bad situations (here communication channels). But this is only true if no alternatives are available. Again, lack of scientific knowledge generally about psychological response to communication channels into computers make it difficult to say anything meaningful about the trade-offs.

7.14 The Real Time Problem

Real time is characterized by at least two criteria. First is the amount of time to process a second of speech. Real time requires that it take one second or less, without too much variance. Second is the delay before the interpretation is available. Real time requires that it be available immediately after the utterance is finished. For speech this is generally the sentence, which amounts to a few seconds of speech. To equal human performance, it should respond with no delay for trivial utterances, and perhaps a few seconds delay for questions that require intellectual activity on the part of the human being. Thus, a system has the right to wait until most of the context has arrived before obtaining the final interpretation. But it cannot wait several minutes until, say, numerous utterances have piled up.

It seems important to accept both these real time criteria as design goals. The usefulness of speech input to a machine depends strongly on its being real time. Many of its advantages disappear if the response to trivial statements is too sluggish. And the data rate advantages disappear almost by definition.

The real time problem strongly interacts with both the amount of processing and of memory available (Sections 7.15, 7.16). However, the only safe view is that in the short run (3-5 years) more powerful general purpose equipment can only convert approximately real time performance into real time performance. It cannot make a silk purse out of a combinatorial sow's ear.

7.15 The Processing Power Problem

An axiom underlying the study was the availability of processing rates substantially faster than existing machines (barring some of those that one would not normally think of as I/O processors, e.g., Illiac-IV). Thus, one should design for the availability of extra power. This is compatible with one of the real time concerns, namely that some set of time constraints be accepted to force concern with the amounts of processing devoted to various parts of the recognition process. It is somewhat at variance with the other real time goal, since one wants to communicate verbally with precursors to the eventual system, and it will be a shame if this is too far off real time, though a factor of five down can probably be tolerated.

We have nothing sensible to say about the actual amount of computation needed, except that it will require substantial amounts of power, and that one should look toward a system on new hardware. Too much depends on the exact nature of parameter extraction processes and of search and match operations.

The extraction of parameters, which involves operations akin to taking transforms, constitutes one place where specialized hardware might make a substantial difference. Specialized hardware is associated with some recent proposals for new parametric representations of the speech wave (the ASCON parameters, introduced by Culler, 1969). We have more in mind -- devices to do variations on the Fast Fourier Transform. The exact nature of such hardware, and whether it is really appropriate, are matters that it is not necessary to settle now. The major concerns seem to us those of the total system and its organization, and not the efficiency of a particular component.

However, this undoubtedly needs serious attention if an effort of major proportions gets underway.

7.16 The Memory Problem

We separate this from the processing problem only to emphasize that one could have plenty of millions of instructions per second and still have problems if one had a 1000 word dictionary with 1000 words of data per lexical item (e.g., many samples) and needed to access the memory a few hundred times in a half a second. The amounts of data lie rather close to the boundary between fast (e.g., core) memory and slow (e.g., disc) memory. As the discussions of memory organization have emphasized, random accessing requirements may pose very serious problems for sequential stores, but not at all for core or film stores. The advent of other stores, such as optical discs, could change these considerations radically. In fact, the memory structure could have a major influence on the entire logical structure of the processing.

7.17 The Systems Organization Problem

Throughout the report we have emphasized the problem of system organization. It has many aspects. First is what information is in the system and how it is represented. We have argued that an organization of levels is mandatory, though more flexible organizations are conceivable. Second is what communication can occur between levels -- e.g., feedback and feed-forward. Implicit in this is a strategy for handling errors. Third is the control structure that permits this communication. The system must have its own scheme of multi-processing or parallel processing structure. The memory structures are too large to permit a simple single-processor-single-random-memory organization. All other forms raise formidable software system problems, where right answers are not easily found. Fourth is the collection of search, match and processing programs. These, of course, are concerned with individual aspects of the total effort, e.g., how to access the lexicon. But which programs are necessary, and with what sophistication and efficiency, depends intimately on the total system organization and computational strategy. It cannot be divorced from the general problem.

The ability eventually to put together a speech-understanding system of realistic dimensions depends heavily on a great deal of development of all these aspects -- of the possibilities for trade-off and how to achieve them. It will not do simply to work on the pieces piecemeal (to coin an alliteration) and then to put together a total system on the basis of a paper design, embodying one set of untested notions for how to handle the total system trade-offs. There must be as much experimentation and development here as in the parts that seem more properly speech oriented.

Unfortunately, research on systems organization cannot go on in the abstract. It is not possible to define any of the issues so they can be attacked effectively outside of specific speech-understanding systems. There is a slight paradox here that reflects the current imperfect state of computer science. Most of these questions have counterparts in other large systems -- questions of file organization and search, of heuristics to avoid looking at low probability material, of control structures, of efficient handling of confidence indicators, of good matching algorithms. Work on these problems in other contexts will help in corresponding work on speech systems and vice versa. Yet, it is not safe to pose these problems in the abstract, or in the context of other fields, as a way of getting the job done for speech-understanding systems. The probabilities of going seriously astray are too great. Rather, substantial amounts of experimentation with total systems must occur prior to the design of a final system.

7.18 The Cost Problem

Ultimately a voice terminal has to be competitive with other channels, except those special applications where voice has distinct qualitative advantages (e.g., hand-free mobile operation). Low speed terminals now cost of the order of a thousand dollars at the terminal and almost nothing internally. A speech system will have these costs reversed: almost nothing at the terminal (a telephone handset at best, but a microphone at worst), but thousands of dollars per terminal internally. For the amount of processing that is involved, even under optimistic circumstances, the cost is very large unless it can be time shared among many inputs.

We have not felt it necessary to attempt these sorts of cost analyses, not does the necessary information exist to do so. Some of the same statistics on multiple users applies here as to multiple teletype users. However, not only is the data rate higher, but speech users will have a tendency to surround meaningful talk with fragmentary utterances. Even if these do not cause excessive additional confusions, they will cause computation to eliminate them. Some of our tasks control this more than others. The point is to caution against the asumption that N voice channels will have the same leisurely statistics that N teletype channels have.

7.19 The Completion Date Problem

Three years, the time initially targeted for the system, is not enough. To produce a system by then requires that everything be in hand scientifically and organizationally. In fact, some directed research is required for an interesting system. Though the critical items have a reasonable chance of success with rather short time scales, their results should be in hand to permit development to proceed.

More important is the requirement for cycles of experience with the total system organization. To produce a system in three years requires immediate commitment to a system organization, including representations and the mechanisms for intercommunication. This allows efforts on components to proceed in parallel, which then come together for assembly and testing at the end. But this is just what will fail, since the systems organization is a major part of the problem.

We do not start from scratch on experience, of course. Though many of the earlier systems of the Fifties were too simple to count as experience in system organization, the system of Vicens and Reddy can be considered relevant. Successor iterations of this system are in progress (at CMU and Stanford) and will provide another cycle of experience within another year. Still, it seems necessary to obtain at least one more cycle of experience and research, with systems that are relatively independent of this line of work, utilizing different design choices and focussing on different aspects.

Five years seems an appropriate time scale, in the light of the above requirements. There are additional reasons for moving beyond three years, but they belong to the next section.

7.20 A Target System

The comments about each of the nineteen problems adds up to the two major conclusions of the study. First, the initial specification, on the left side of Figure 1.1, is too ambitious given the state of the current art. Second, gathering together the specifics of our comments yields a modified specification, on the right side of Figure 1.1. This specification seems to offer a reasonable chance of attainment, if pursued with appropriate motivation and funding. These latter issues belong to the next and final section.

8. WAYS AND MEANS

We have arrived at the following position: The initial specifications are too ambitious, even within a five year period, but we have replaced them with another set of specifications that seems possible of accomplishment. We accepted some items which are beyond the art today, but seem necessary for a total system to constitute a major step in the right direction. Chief among these were continuous speech, many speakers, and real time. We specified a usefully large vocabulary (1000 words), though its effects may be primarily through the real time problem.

We eliminated items that, though not necessarily harder, seemed less critical to an appropriate forward step: the telephone, speaker variation by age, sex and dialect, and the use of general English-like languages. For each, specific areas of ignorance were identified, and great risk attaches to multiplying the number of independent major successes required to obtain a system. Some of these (the telephone and general syntax) seem to be the next steps beyond the existing specification. Immediate research on them is appropriate. With positive enough results from these researches, specifications for the target five-year system could be upgraded. Others (e.g., speaker variation by age, sex and dialect) seem postponable.

This section discusses how the five year target system might be achieved. There is no one way, and any plan must respond ultimately to available technical resources and funds. Nevertheless, the easiest way to present our ideas and opinions is to lay out a specific plan. We did in fact consider a number of alternative arrangements, which are presented briefly in Appendix 12. But only one variation received sufficient attention in the light of the actual target system to be presented in detail.

8.1 The Plan

Figure 8.1 lays out a plan for achieving the target system in five years. It starts by initiating directed research in three areas: (1) phonological and acoustic-phonetic rules; (2) the syntactic interface; and (3) recognition experiments on the telephone. The first two are more substantial efforts than the last, in terms of manpower and scientific involvement. A summer institute occurs in 1972 to launch the appropriate research and spread interest in it throughout the scientific community.

Also in the first year potential candidates for putting together the target system are selected and funded. These organizations require sufficient funding to acquire appropriate people and to construct an experimental speech-understanding system. This latter, to be created during the first two years of effort, is both a vehicle for research on system organization problems and real time problems, and a major input to the selection of final contractor(s).

A major structural feature of the plan is a decision point at the end of the second year. At this point it will be decided whether or not to attempt a three year development effort for a speech-understanding system, and, if so, with how many parallel efforts. Substantial new information will be available at this time, so that the decision will be a real one. These inputs, discussed in Section 8.4, are:

(1) Results of phonological and acoustic-phonetic research.

(2) Results of syntactic interface research.

(3) Results of studies on the effect of telephone communication.

(4) The preparedness of the candidate contractors, including:

  (a) Experimental total speech-understanding systems.

  (b) A detailed analysis of the tasks to be performed by the system, and the range of mechanisms to be used.

  (c) A decision on the parametric representation of speech to be used, supported by adequate technical analysis.

We expect, of course, that two years' work will have produced additional knowledge to help estimate the situation. However, we are banking only on the items above.

The possible outcomes of the decision point are:

(1) To go ahead with a single contractor.

(2) To go ahead with several contractors, presumably with variation of systems and tasks sufficient to warrant multiple efforts.

(3) To delay the decision, since no available contractor has the requisite capabilities to produce the system, even though the scientific indications are positive.

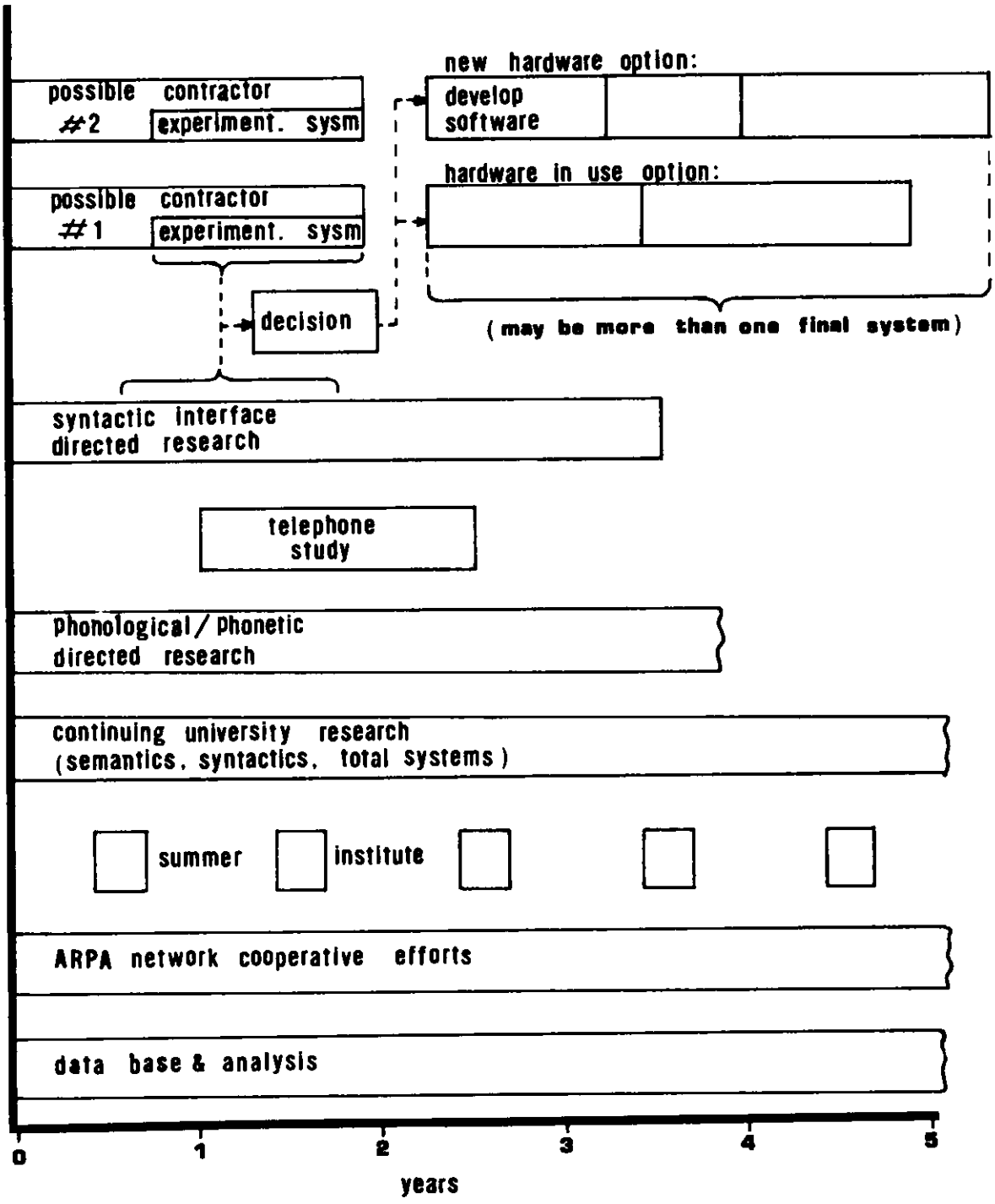


Figure 8.1 PLAN FOR DEVELOPMENT

(4) Not to go ahead, since the scientific results (including questions on the total system) are not encouraging:

Continue research on the difficulties.

Terminate the research program.

Assuming a positive outcome, an important decision about hardware facilities would also occur at this point. The substance of that decision does not concern us here; the proposals would undoubtedly contain explicit specification for the hardware to be used. However, if the decision is made to use brand new hardware, then at least another year should be inserted into the time scale. On a newly manufactured computer system the software and operating systems will be in rudimentary operational shape (though no doubt in excellent conceptual shape). The organization will put a substantial amount of effort into augmenting the software systems to bring them to the state of convenience commonly expected on the prior system.

The position just enunciated is not a bias for living with old systems. Speech-understanding systems must eventually move to more powerful hardware. But we do not underestimate the necessity for adequate software, especially in constructing a total system of the magnitude of the target specifications. Thus, two branches occur after a positive decision.

The development systems will proceed like a normal development contract to a single group. It makes little sense to guess now at intermediate stages for a target system to go through. Each contractor is tooled up with a relatively experienced functioning organization at the start of the three years. The initial version of the total system should come into being relatively early in the three years. We have marked this at the beginning of the second year, though it may be a little later. This permits substantial adaptation and redesign of the total system, so that the final version at the end of three years may contain little of the initial code.

The target system is not the end of the line. We included early research to upgrade the system specification. Even if these early attempts fail, it is entirely likely that additional work will eventually bring to fruition adequate grammars for errorful input and semantic systems for question answering that are appropriate for speech-understanding systems. Continued work on phonological and acoustic-phonetic rules will be highly valuable. It will begin to exhaust the backlog of studies that currently exist and will be coupled with new investigations. By that time the amount of speech data processed and examined analytically will exceed the amount done throughout the past. In short, research must continue throughout the five years, though its level is not high with respect to the other expenses.

At the end of the program there is a relatively heavy concentration on one (or a few) contractors. However, we do not think the program will work without a general commitment by a number of investigators. This holds especially in the early phases, but remains true throughout. Several mechanisms are worth serious consideration both to increase the general involvement and to help keep it focussed. The summer institute is the first one. The existence of a committee of IPT contractors involved in speech-understanding systems is a second. The use of the ARPA network is a third. The continuation of university-based research is a fourth. And the existence of some means of obtaining high quality data for test and comparison of systems is a fifth. These are discussed in Section 8.3.

8.2 Specification of Initial Research

Four areas of research are appropriate for immediate attention. Examples of each have occurred in the analysis section, so we concentrate here on programmatic statements of what research is to be done.

Phonological and acoustic-phonetic rules. Initially, the task is to take existing knowledge and convert it into a form that is useful for recognition. The data bears on two levels of system, and, as it occurs in the scientific literature, is inverse to the form most needed for recognition.

The phonological rules are well typified by the work of Chomsky and Halle (1968), but also by much classical work in linguistics (Bloomfield, 1933; Fries, 1952; Hill, 1958; Hockett, 1958; Trager and Smith, 1957). From a phonemic representation of the lexical item they derive a phonetic representation. In almost all modern work (since Jakobson, Fant, and Halle, 1963) this is represented by some system of distinctive features. The rules thus serve to produce a correct phonetic representation, which can be related to the parameters of speech.

One wants several things. First, one would like a handbook of all rules in a form that makes it easy to understand the rules from the viewpoint of recognition as opposed to production. Each recognition effort should not have to engage in its own literature research to dig up each minor rule. An indication of the actual evidence for these rules would be useful. Second, one wants versions of the inverse rules: those that say that such and such combinations of distinctive features are prohibited, because of phonological rules, or that such and such a pattern

of distinctive features indicates the occurrence of certain phonemic sequences. The form of such rules, and the notation in which they should be cast, are matters of some moment (and beyond our present endeavor).

The acoustic-phonetic rules, as they now exist in the literature (e.g., Lehiste, 1967), relate the parametric representation of the speech signal to the phonetic context in which it occurs. Again, these need to be cast into a handbook form using some uniform representation. And, as with the phonological rules, some form of inverse rules need to be derived, i.e., those that go from features of the parametric representation to selections and exclusions on the phonetic representation.

These bodies of knowledge must be generated in a general form, not embedded in a particular recognition algorithm. Of course, by the act of choosing a notation, and selecting forms for the inverse laws, some bias is created. But the risk seems negligible compared to the gain from having a body of rules available for use in many forms of algorithms.

The rules, as they now exist, are not only scattered in the literature, but in the heads of the scientists who are (or have recently been) working at the main centers of such research. As usual in scientific work, there are only a few such places. Here the set mostly includes the groups around Fant at the Swedish Royal Institute of Technology, Stevens and Halle at MIT, Lehiste at Ohio State, Ladefoged at UCLA, the group at the Haskins Laboratory, the group at Bell Laboratories, and the group at the Speech Communications Lab at Santa Barbara. The success of this part of the plan is critically dependent on inducing a significant interest in the problem at some of these places.

The proposed endeavor involves theory as well as compilation. It also involves substantial systematic testing. Not all the work in the literature is empirically complete and sound. Furthermore, inverting rules can produce inappropriate generalization and must be tested experimentally.

The total task involves substantial numbers of man years. Yet, enough can be done in two years to testify to its usefulness and to permit estimation about the effort required to bring forth additional results.

Syntactic interface. The analysis section made clear the kinds of interface needed between the syntactic level and the levels below:

(1) How to parse sentences with errors in the terminal symbols.

(2) How to parse sentences that are seriously ungrammatical, involving fragments, repetitions, etc., as in natural speech.

(3) How to use confidence symbols attached to terminals in parsing.

(4) How to parse when the order in which the terminals are presented is determined by the lower levels.

(5) How to parse when the information given consists of phonetic features rather than full words.

(6) How to use information about pauses, stress, and intonation in parsing.

(7) How to provide subsets of lexical items that might limit the possibilities at a point in the phonemic representation.

All the questions are approachable, and some possibilities were explored in the analysis section in terms of a particular grammar (that of Woods). However, initial attempts to solve these problems are rather crude, i.e., they retreat to extensive generating and testing. Modern grammars are characterized by efficient parsing algorithms, which are well adapted to their basic assumptions of a correct and complete input string, available at the beginning of the parse. The effort used in parsing is important, especially since it must increase in any event with errorful and fragmentary prose.

Thus, answers of various degrees of sophistication can be produced for all the questions posed above. Further, all the questions (especially the first four) have some intrinsic interest to those who work on grammars. In short, we believe that if these questions were motivated properly, the relevant community would pick them up and work on them.

Unlike the work on phonological and acoustic-phonetic rules, a widespread group of people in computer science are competent to work on these syntactic problems. Furthermore, all that is required, beside the requisite expertise, is adequate computer facilities.

The telephone system. We have already stated our unwillingness to consider the ordinary telephone as a suitable communication system. This position is mostly an assessment of ignorance, although we know enough to be cautious. However, it seems possible to determine whether or not the telephone can be viewed simply as another noise source, whose degradation can be measured by a few parameters and which can be included in a system with an expected degradation of X% in recognition capability. The alternative

(negative) view is that the ordinary telephone in the field is sufficiently variable and has sufficiently perverse noise so that it becomes a special problem, one that belongs further down the scale of priorities.

The required study compares the effectiveness of relevant algorithms on known samples of speech. These analyses start at the lowest levels of representation, working up until no telephone-specific effects are discernable. Enough telephone conversations must be used to sample the range of actual conditions as a function of relevant parameters of the telephone system (e.g., line length, number of switching centers, carbon button microphone variation, etc.)

Such a study is relatively straightforward, given that the recognition algorithms and test data have all been generated for other needs. Thus, initiation might not occur immediately, but certainly such studies could be completed before the two-year decision point. Specific algorithms must be picked, which necessarily introduces some particularity into the results. Still, such information will be extremely valuable, and should remove the state of ignorance which is the study group's current plight.

Real time processing and system organization. The final system will consist of many cooperating parts, operating in an environment where memory and processors must be shared in order to get the total job done. Homogeneous organizations seem too radical a departure for the presented target system. Thus, the expected organization consists of a collection of representations with systems for intercommunication between them, possibly by means of a multi-processing system.

Research is badly needed both on the overall organization and on the specific schemes for working with large files and expensive matches. Given the present state of computer science, these problems cannot be attacked except in the context of specific systems. They cannot be abstracted. Each effort at a total speech-understanding system must also engage in deliberate study of these aspects of the total system problem.

It is not enough that each total system implicitly contain an overall design. If attention is not directed to the issues, analysing both the possibilities at design time and the costs and benefits after the system comes up, then little informed experience will exist by the time the final system is designed. Man's ability to design large systems foolishly, because all he has is foresight, is well documented in existing large software systems. It is absolutely necessary to obtain at least one (and hopefully two) cycles of design-construct-analysis on the system organization issues.

8.3 Cooperative Endeavor, Control and Public Information

The effort to get a significant speech-understanding system in five years must blend both science and development. Our plan clearly has elements of both. We have given some thought to the mechanisms by which a scientific community can be induced to work intensively on this problem. The basic ingredient, of course, is that the problem be scientifically exciting and lead beyond itself. To us, speech recognition is such a scientific domain; also, as we have laid the problem out here, it has none of the unfortunate detractions that seem often to accompany work on systems to perform human perceptual and intellectual functions. We are not so sanguine to think no one will view the effort from such a detracting viewpoint. The evidence is all to the contrary (Pierce, 1969). Nevertheless, speech recognition, as described herein, seems to us both scientifically respectable and intellectually engaging.

More than this is required, however. The problems must become known to the scientific community. Effort must be applied to the solution to this problem, rather than some other problem (for science has a certain random walk character). The feedback of results must be intensive and accurate, to avoid error and accelerate progress. The collection of mechanisms below are directed toward these ends.

Summer Institute. We propose that a summer institute be held as soon as practical after the beginning of the program. This would help define the various research problems, make them known to the relevant community, create strong initial forward momentum, and produce agreement on numerous aspects (e.g., notations for acoustic-phonetic rules). In short, it would take a step to create the relevant scientific subcommunity. Necessary to such an insitute is a substantial speech processing facility, so that a good deal of minor experimenting and playing around could go on. Exactly how long, where, who, etc., is a task for another group to deal with. But some event of this type would be beneficial.

Whether there should be summer institutes of various characters in other years, we leave open.

Steering Committee. A way is needed to provide continuing guidance, to foster cooperation, to avoid inappropriate duplication, to communicate informal results and to force agreements when technically required. The pace of the five years is too fast simply to let research contracts at the beginning, to come to fruition or die as they will.

We considered several mechanisms, varying from running the whole by a single contractor, to having a meta-group with funds concerned with the whole, to permitting a laissez-faire organization.

With the current plan, which delays the actual development contract until about the middle of the course, a simple steering committee structure seems best.

Each IPT contractor working on the speech-understanding systems effort would have a member on the committee. Others might also be invited to join with full privileges, even though their funding came from elsewhere, if they were involved in the total effort functionally and were internally committed to a cooperative endeavor. Each case would be handled on its merits.

The committee would meet at least quarterly for the first two years, possibly more often right at the start. It would be chaired by the staff person in IPT responsible for the program. Thus, formally the committee would be as an advisory committee to IPT on the conduct of the speech-understanding-system program (or whatever its official name became). Whatever powers of enforcement the committee would have over its own members would therefore derive directly from the common role all members would have as contractors to IPT.

The functions of the committee would be to keep the program under continual review, to see ahead to additional necessary research, to avoid duplication, to help arrange for one contractor to use the results of another, and so on. It would be concerned with how to evaluate systems being developed, how to obtain useful test data, why A's system seemed to work better than B's on particular test data, etc. Presumably, energy devoted to the committee by the members would be energy well spent in the direct prosecution of their respective contracts. Presumably, the issues dealt with would be mostly technical. In fact, it might well conduct meetings in an essentially open fashion so that others interested in the problem could participate in the proceedings.

This steering or advisory committee seems absolutely essential for the first couple of years, through the major decision point. Its role could then diminish somewhat in importance, although it might well continue to be moderately effective under a somewhat lower head of steam for several additional years.

The ARPA Network. A novel aspect of the current situation is the emergence of the ARPA network. This offers an enticing set of possibilities for a cooperative scientific-development effort, such as the one under consideration. It offers the possiblity that various members can use each other's programs, hardware, and data, permitting more rapid and effective experimentation at higher levels of system organization.

The early state of development of the network prohibited incorporating it into the plan in any central way. Nevertheless, we think it important (and worth resources) to promote active cooperation among the community of contractors working on a speech-understanding system via the network. Given the necessarily imperfect nature of the network initially, this requires devoting talent directly to network use. Otherwise, real conflicts will occur with other more productive use of contractors' energies, especially those whose programs would initially be used by others, and who may be most capable of moving ahead substantively.

University Based Research. The major share of current IPT research in speech recognition systems already occurs at universities. They are not appropriate, generally, as candidates for carrying out development efforts. Universities will, of course, be involved heavily in carrying out the early directed research. But it is also appropriate to continue a substantial university level of research into speech-understanding systems more generally.* The arguments are the standard ones, but no less effective for that. Many, if not most, conceptual advances come from universities. (Though speech is a somewhat unique partial exception, due to the presence of Bell and Haskins Laboratories, both long term and effective participants in speech research.) We have laid out a seemingly tidy research plan. But, in fact, to obtain the target system, the level of activity on speech-understanding systems throughout the country should be raised by a substantial increment to encourage new developments (not all IPT supported, of course). It is quite possible for new systems to emerge in university environments sometime during the five years that will be serious contenders for a speech-understanding system.

Public Data and Public Analysis. A major instrument for progress on speech-understanding systems will be good data of suitable variety, prepared so that it is possible to relate how different systems and algorithms process it. Claims will be made about a wide variety of systems and subsystems over a wide variety of communication situations. If the claims are not made against a background of publicly available high quality data of known structure, it will never be possible to understand the claims or their basis. The issue is not one primarily of assigning credit, but of making progress by understanding success and failure.

---

* We re-emphasize the point made in the preface: we are talking about universities generally, not necessarily those with current IPT contracts. An extensive effort, such as that described in this report, would undoubtedly involve many new IPT contracts; and existing IPT contractors would not necessarily be involved.

We would extend the concern for good data in three directions, without, hopefully, diluting the concern thereby. First is adequate task description. We consider the move made in this report to consider highly specific, well specified task environments to be a step in the right direction, though of course only a preliminary one. Taking extensive behavior data over suitable ranges of variation is a most profitable activity.

Second is instrumenting the systems (both hardware and software) and taking appropriate measurements. Measures of total performance (e.g., percent semantic errors), though absolutely essential, are almost useless in pinning down the causes of performance. Almost all papers in the computer science literature on large systems are deficient in the measurements taken. Speech-understanding systems have meaningful internal interfaces at which measurements can be taken. Furthermore, errors can be traced to the algorithms or data that caused them, rather than being lumped together in summary statistics. Measurement is very difficult to do after the fact. Plans for instrumentation and measurement on a routine basis must be part of the systems design.

Last is modeling the system. The class of speech-understanding systems lends itself to the construction of operations research type models that attempt to parcel out the total performance of the system among the mechanisms. Very little of this has been done for complex software systems of any kind. However, considerable opportunity exists to do it meaningfully for speech-understanding systems. Our efforts in Appendix 10 exhibit the spirit in which we think such modeling can be done (though preliminary in its results).

These three directions in which the proposed system can be measured and analyzed constitute a proposal for a highly rational development of the target system. Partly, we propose it because the systems lend themselves to such analysis. But also, such public analysis will accelerate the development of the target system appreciably.

A major function of the steering committee would be to give these notions operational form. But resources must be devoted to these activities directly. Some other institutional mechanism must be used, though we have no definite preferences -- whether by separate organization, or by people attached to each organization, etc.

8.4 Requirements for the Contractors Developing the Target System

We wish to lay down specific requirements for a potential contractor to take on the job of producing the target system in three years. In part these requirements assure that the chances of success can be evaluated. We leave open how the evaluation of the contractors (and the state of the first two years' research) is accomplished. We describe here the set of requirements and their rationale.

Operating Total Speech-Understanding System. Two main purposes are to be served. First, it assures that the candidate is tooled up and ready to go, especially that key people are on board, which is a most time consuming aspect of tooling up. Second, it demonstrates that the candidate has some total systems capability. This is important, and the only way of demonstrating it is to show the sorts of total systems one has put together. As already mentioned, this requirement also serves the goal of getting research done on the system organization problem, but that is not the relevant consideration for this subsection.

Having a running total system does not imply, of course, that the candidate has a version of the target system. He must have put together and made to work a system with a number of levels of representation in it. This undoubtedly will be an entirely experimental system, and may contain components that are parts of the systems of others. The Vicens-Reddy program shows that such systems can be put together. A second time around it can be done simply as a serious exercise. The candidate, no doubt, will use the exercise to explore his own notions about how to organize such a system.

Settled Parametric Representation. Currently, several parametric representations are available: The time domain signal itself; a set of zero crossings and amplitudes; a set of filters; the ASCON parameters; and the parameters derived from the articulatory representation of the speech signal. Preferences certainly exist among these. An obvious feature that differentiates the last one is its close relation to the acoustic-phonetic laws, it being the only representation in which they currently have a natural representation.

We are not prepared to specify which parametric representation is appropriate, though members of the study group have preferences of their own. It is appropriate that a candidate have determined which parametric representation it shall use and be able to give an adequate technical defense of this choice. Leaving this decision go until later jeopardizes too much, in leaving some basic structure up in the air until too late in the game.

Adequate, Detailed Task Description. Our earlier remarks about the virtues of good descriptions of task environments is applied here. Given such a description, including protocols with audio tapes, grammars, vocabularies, etc., reasonable assessments are possible about the proposal. Without this, an assessment is shooting in the dark.

Adequate, Detailed System Design. Letting the decision on contractor go until the end of the second year makes it possible to expect a detailed design of the proposed system. Our bias is that such a design can include preliminary before-the-fact performance analyses which will prove extremely revealing about the possibilities of success. On both this and the previous requirement, it is not our intent to stop modification of the proposal in the light of future developments. Such changes should take place against an appropriate background analysis.

Adequate Instrumentation and Performance Analysis. This requirement reflects our assessment, stated above, that only by obtaining adequate measurement in adequately defined environments will appropriate feedback occur. As noted, the instrumentation, especially, must be planned into the basic system or it simply will not occur.

Proposal for the Hardware to be Used. If there is to be new hardware (as may well be), then the proposal should contain, as well, an adequate plan for the development of software on the new machine.

A1. HISTORY AND STAFFING OF THE STUDY GROUP

On March 30 to April 1, 1970, a meeting was held in Pittsburgh at Carnegie-Mellon University to discuss the question of the feasibility of a speech-understanding system and to determine whether a study group should be set up to examine the question in detail. This meeting was initiated by the Information Processing Technology Branch of ARPA following discussions held at the yearly meeting of the IPT contractors held in New Orleans, in January, 1970, plus numerous informal discussions. As the report notes, IPT has been supporting a modest amount of work on speech processing related to computers (both synthesis and recognition). Some new indications of progress along with the general advancement of computer technology provided a sense that technical assessment was appropriate.

Almost all the IPT contractors with an interest in speech were represented at the Pittsburgh meeting. General assessments of the art at the different levels of the system were briefly presented and discussed. It was agreed generally that an investigation of feasibility was appropriate, though there was considerable doubt about the attainability of the initial specifications, as they were outlined.

The assumptions that become integral to the study group's thinking were already in evidence at the initial meeting: (1) it is necessary to consider a total system with major support from the syntax and semantics; (2) there does not exist a large technical literature to be uncovered and assessed; (3) what was needed was investigation of the propects of a total system. It followed that any study should be of short duration.

The study group was formed at the Pittsburgh meeting, with its personnel all drawn from those present. A major factor operating in the (self) selection of the group was the ability to make the necessary time available on almost instantaneous notice (many who would have participated already had firm summer plans). The study group, as it emerged, contained a substantial amount of experience in the area of computer oriented speech recognition systems, in the sciences of speech, and in work in syntactic and semantic processing. Since the expertise of this group is of legitimate concern in assessing the report, we give a brief account of the relevant experience of each member at the end of the appendix.

The study group met again on May 26-27 in Boston at Bolt, Beranek and Newman. It had been decided at the first meeting that an appropriate tactic for analysis was to consider in detail some highly specific tasks. The Boston meeting was devoted to selecting specific tasks for a speech-understanding system, to be used for the remainder of the study. The tasks discussed in the body of this report emerged from this meeting.

The final meeting of the group was held on July 26-28 in Santa Monica at the System Development Corporation. By this time a number of partial position papers, lists of questions, etc., had been accumulated. The meeting settled on the recommendations as presented here and on the essential content of a report. The drafting of this final report was undertaken after the meeting.

* * *

Biographies

Dr. Allen Newell is University Professor at Carnegie-Mellon University. He has made numerous major contributions to artificial intelligence and to the modeling of human problem solving.

Mr. Jeffrey Barnett is a computer systems specialist with the Advanced Development Department of System Development Corporation. He has worked in programming language development and list processing and is currently project leader of SDC's Voice I-O Project.

Mr. James Forgie is Associate Leader of the Computer Systems Group at Lincoln Laboratory. He was one of the first researchers to apply computer techniques to phoneme extraction and recognition.

Dr. Cordell Green is Research and Development Program Manager of the Information Processing Techniques Office of the Advanced Research Projects Agency. He has worked in the theorem proving and question-answering areas of computer science.

Dr. Dennis Klatt is a Research Associate in the Speech Communications Group of the Research Lab of Electronics at MIT. He has developed techniques for extracting phonetic properties of speech and has worked on speech synthesis by rule.

Dr. J.C.R. Licklider is a Professor of Electrical Engineering at MIT and is the Director of Project MAC. He has made significant contributions in the areas of speech communication systems, speech compression, psycho-acoustics and speech recognition, and more recently in the areas of time-sharing and man-computer symbiosis.

Dr. John Munson is a Senior Research Physicist in the Artificial Intelligence Group of the Information Science Laboratory of Stanford Research Institute. He has worked on speech recognition and automatic recognition of hand-printed characters utilizing contextual constraints.

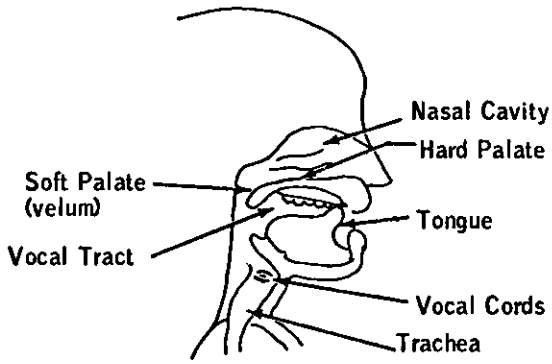
Dr. D. R. Reddy is an Associate Professor of Computer Science at Carnegie-Mellon University. He has worked in the areas of voice and visual input to computers, and directed the development of the first demonstrable real-time connected speech understanding system.

Dr. William A. Woods is a Senior Scientist in the Artificial Intelligence Department of Bolt, Beranek and Newman. He has developed parsing and semantic interpretation techniques for handling English communication with computers.

A2. INTRODUCTION TO SPEECH

A comprehensive survey on speech recognition research can be found in Lindgren (1965) and Hyde (1968). Most of the relevant material on speech analysis, synthesis and perception is discussed by Flanagan (1965). There are numerous textbooks on phonetics and linguistics (Bloomfield, 1933; Trager and Smith, 1957; Hill, 1958; Hockett, 1958; Chomsky and Halle, 1968). Thus, in this appendix, we will restrict ourselves to some basic material on speech.

Sounds as a listener perceives them are the result of compression and rarefaction of the surrounding air impinging on the ear drum. Such changes in air pressure may be caused by the vibrations of a string, a surface or a column of air. The human vocal apparatus is one such complex assembly for producing sounds, both tones and noises.



The vocal tract, whose cross-sectional area can be modified by the movement of the lips, jaw, tongue, and velum, provides the main column of air, which may be set to vibration at its natural frequencies by a suitable energy source. The nasal tract provides an auxiliary column of air which may be coupled or uncoupled with the vocal tract by appropriate movement of the velum (soft palate). During expiration, the moving air provides the necessary source of energy for speech production. This moving stream of air is acted upon by various parts of the vocal mechanism to create various acoustic disturbances which are perceived by the listener.

In the case of voiced sounds the air stream is permitted to excape in quasi-periodic pulses by the vibratory action of the vocal cords. This in turn sets the acoustic system above the vocal cords vibrating at its natural frequencies. These resonant frequencies of energy concentration are known as formant frequencies. They are useful in characterizing the vocal tract configuration although there exists no one-one relationship between the vocal tract configuration and the formant frequencies.

In the case of unvoiced sounds the vocal cords are relaxed and partially open. Turbulent flow of air is created either due to some point of stricture in the vocal tract or due to the abrupt release of the pressure built up at some point of closure in the tract.

Since most of the vocal organs can be moved in many ways by volitional muscular activity and since it takes but minute alterations in the vocal organs to produce distinguishable varieties of sounds, a human is capable of originating an enormous variety of distinct tones and noises. Of this vast potential of articulations each language employs only a rather restricted number of classes or articulations. Every such class is called a phoneme. It is appropriate to quote Flanagan (1965) on this topic.

> "To be a practical medium for the transmissions of information, a language must be susceptible of description by a finite number of distinguishable, mutually exclusive sounds. That is, the language must be representable in terms of basic linguistic units which have the property that if one replaces another in an utterance, the meaning is changed. The acoustic manifestations of a basic unit may be subject to great variation. All such variations, however -- when heard by a listener skilled in the language -- call up the same linguistic element. This basic linguistic element is called a phoneme (BLOCH and TRAGER). Its often-manifold distinguishable variations are called allophones."

Of all the different kinds of articulations used by the humans we shall restrict ourselves to those used in English speech. Of course, any general speech recognition system must possess greater discriminatory ability to correctly recognize those sounds which are not phonemically distinguishable in English. Pitch inflections (Chinese), whispered vowels (Japanese) and vocal clicks (South African Hottentots) are some examples of speech sounds which are phonemic in other languages and not in English.

Phoneticians usually classify speech sounds by specifying the manner and the place of their production. Another approach to phoneme classification was devised by Jakobson, Fant, and Halle (1963) using the distinctive features of the speech sounds.

In the former approach sounds are generally described by the position of the tongue hump along the vocal tract, the degree of constriction, presence or absence of voicing, turbulence due to non-laminar air flow and such features. The vowel sounds are specified by the position of the tongue hump and the degree of constriction of the vocal tract. This configuration of the tract is maintained stable while the vowel phonation occurs. The vowels usually have a higher acoustic power than the consonants resulting from the relative absence of tract constrictions. The vocal tract excitation by the vocal cords contributes most of this power and only negligible amounts due to nasal coupling (except when a vowel is nasalized as in French). The tongue hump position and the degree of constriction of the English vowels of General American dialect are shown in Figure A2.1.

Consonants, unlike vowels, are not exclusively voiced and mouth-radiated from a relatively stable vocal configuration. Presence or absence of voicing, presence or absence of nasal coupling and the short time dynamic motion of the vocal apparatus are useful in classifying the consonants. Fricatives, nasals and semivowels may be uttered as sustained sounds whereas stops and glides depend on the dynamic movements of the vocal apparatus for proper articulation. Nasals, glides and semivowels are always voiced. Fricative consonants are characterized by the noise produced by the turbulent airflow at some point of constriction. Common constrictions are those formed by the tongue behind the teeth (dental), the upper teeth on the lower lip (labio-dental), the tongue to the gum ridge (alveolar), the tongue against the palate (palatal) and partial closure of the vocal cords (glottal). A fricative may be voiced or unvoiced depending on whether the vocal cord excitation is present in conjunction with the noise source or not. Figure A2.1 shows the fricative consonants classified accordingly.

Stop (or explosive) consonants are produced by the sudden release of pressure built up behind some point of complete closure. The explosion and the aspiration of the air and the associated vocal tract dynamics help to characterize the stop consonants. Figure A2.1 shows the classification of the stops according to the point of closure and presence and absence of voicing.

Nasal consonants (m, n, ŋ) are characterized by the complete closure towards the front of the vocal tract and the almost exclusive sound radiation from the nostrils. Figure A2.1 shows nasal consonants classified according to the points of closure.

Glides (w, j) and semi-vowels (r, l) are voiced and mouth radiated sounds. Glides are dynamic sounds which depend on vocal tract movement for proper articulation. Semivowels can be sustained. Figure A2.1 shows their classification according to the place of articulation.

Distinctive Features.

Jakobson, Fant, and Halle in their now classic treatise, "Preliminaries to Speech Analysis" (1963) advance the theory that there exist certain minimal distinctions among phonemes which permit each phoneme to be distinguished from the others. After careful examination of several language structures, they present twelve or so binary choice opposing qualities of sounds (such as voiced vs. unvoiced, nasal vs. oral) called the distinctive features (Figure A2.2). These distinctive features grouped together provide for a unique identification of the phonemes. Some distinctive features and their acoustic correlates are listed below. A more recently revised version of the distinctive features can be found in Chomsky and Halle (1968).

The following acoustic characteristics of distinctive features were given by Halle:

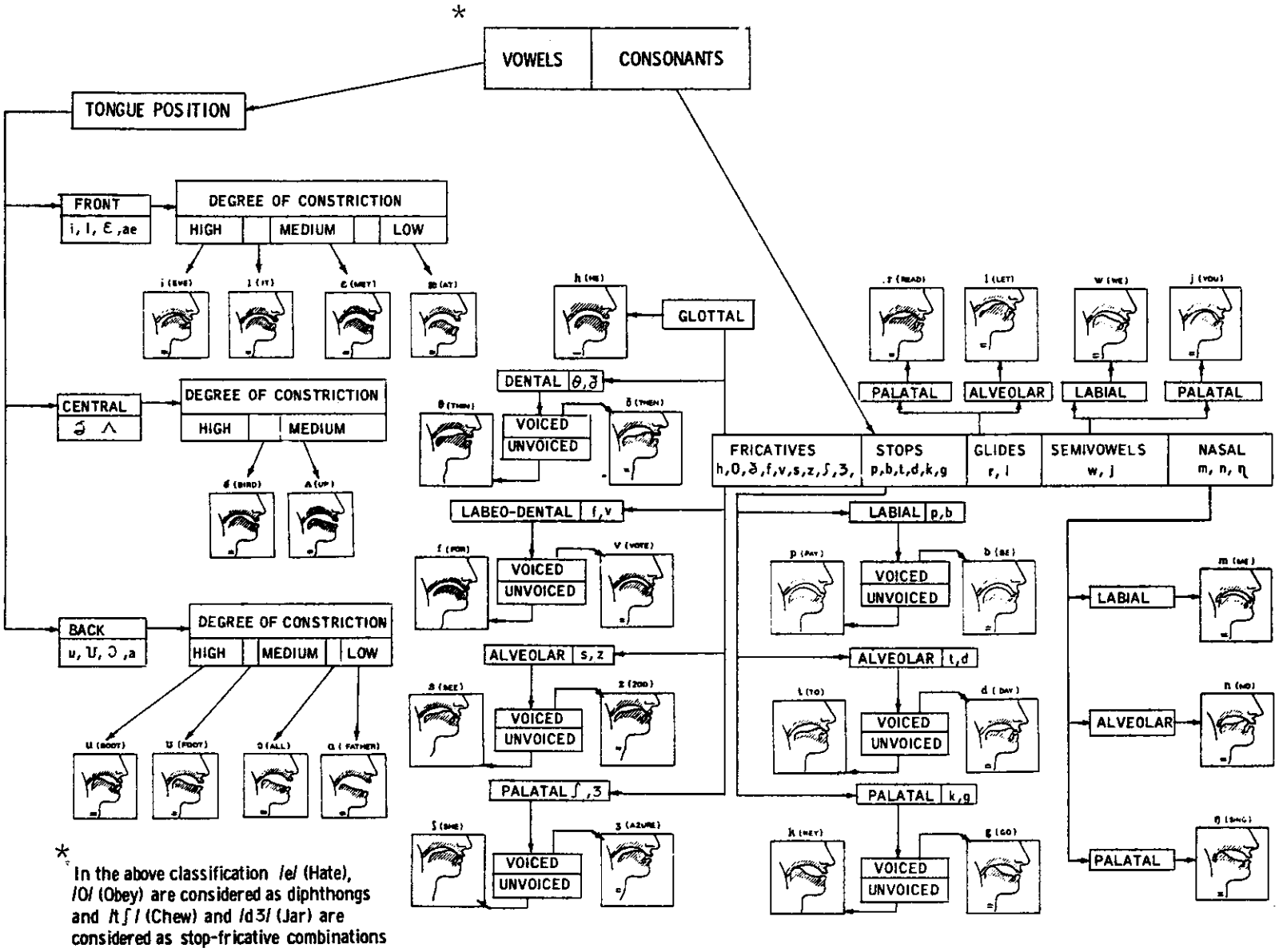
Vocalic/nonvocalic:
Presence vs. absence of a sharply defined formant structure.

Consonant/nonconsonant:
Low vs. high total energy.

Interrupted/continuant:
Silence followed and/or preceded by spread of energy over a wide frequency region (either as a burst or a rapid transition of vowel formants) vs. absence of abrupt transition between sound and such a silence.

Nasal/Oral:
Spreading the available energy over wider (vs. narrower) frequency regions by a reduction in the intensity of certain (primarily the first) formants and introduction of additional (nasal) formants.

Tense/Lax:
Higher vs. lower total energy in conjunction with a greater vs. smaller spread of the energy in the spectrum and in time.



* In the above classification /e/ (Hate), /O/ (Obey) are considered as diphthongs and /tʃ/ (Chew) and /dʒ/ (Jar) are considered as stop-fricative combinations

FIG. A 2.1 - PHONEME CLASSIFICATION
(Figures adapted from Potter Kopp and Green)

DISTINCTIVE FEATURES	PHONEMES																																		
	ɚ	i	I	ɛ	æ	a	ʌ	ɔ	ʊ	u	j	r	w	l	m	n	ŋ	ʃ	s	f	e	ʒ	z	v	ð	tʃ	k	p	t	dʒ	g	b	d	h
1. Sonorant / Nonsonorant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2. Consonant / Nonconsonant	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
3. Continuant / Interrupted																		+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-		
4. Nasal / Oral											-	-	-	-	+	+	+																		
5. Tense / Lax	+	+	-	-	+	+	-	+	-	+								+	+	+	+	-	-	-	-	+	+	+	+	-	-	-	-		
6. Compact / Diffuse	-	-	-	+	+	+	+	+	-	-					-	-	+	+	-	-	-	+	-	-	-	+	+	-	-	+	+	-	-		
7. Grave / Acute	-	-	-	-	-	+	+	+	+	+	-	-	+	+	+	-		-	-	+	+	-	-	+	+			-	-			+	-		
8. Flat / Plain	+	-				-		+			-	+	+	-																					
9. Strident / Mellow																				+	-			+	-	-	-			+	-				

FIG. A 2.2 - DISTINCTIVE FEATURES OF THE PHONEMES OF ENGLISH.
(From Hughes and Hemdal)

Compact/Diffuse:
Higher vs. lower concentration of energy (intensity) in a relatively narrow, central region of the spectrum accompanied by an increase (vs. decrease) of the total energy.

Grave/Acute:
Concentration of energy in the lower (vs. upper) frequencies of the spectrum.

Flat/Plain:
Flat phonemes in contra-distinction to the corresponding plain ones are characterized by a downward shift or weakening of some of their upper frequency components.

Strident/Mellow:
Higher intensity noise vs. lower intensity noise.

A3. DATA ON HUMAN PROCESSING RATES

There is little systematic data on the rate at which humans can utilize various communication channels. Many aspects of the structure of the channel, the task, the knowledge about the message, and the skill and knowledge of the human affect the rates. There are also differences between burst rates and sustained rates for various durations. Furthermore, as in any complex system, the factor that limits the communication rate will vary with the situation, e.g., sometimes it may be a central limitation in forming the message, sometimes a device limitation, as in the motor system of the hands and fingers while writing writing. Given all the above variability, the numbers in Figure 3.2 are simply ball park figures, garnered from a number of disparate sources.

Reading out loud. The attempt to view the human as an information theoretic channel has been the occasion for a number of determinations of reading rates (though not of the spontaneous generation rates). An extensive collection of data can be found in Pierce and Karlin (1957) and in Quastler (1955). A typical one is discussed in Woodworth and Schlossberg (1954, page 508). Data is given for university students (1908 vintage) on reading an interesting novel:

Oral reading:	2.2-4.7 words/sec
Oral reading (try harder):	2.9-6.4 words/sec
Silent reading:	2.5-9.8 words/sec

The rate goes down with technical material and with educational level, as one would expect.

Speaking (spontaneously). Goldman-Eisler (1968) provides some data on speaking spontaneously. The following graph shows the distribution of rates for a single individual during debates, each sample being an utterance from interruption to interruption by another debater.

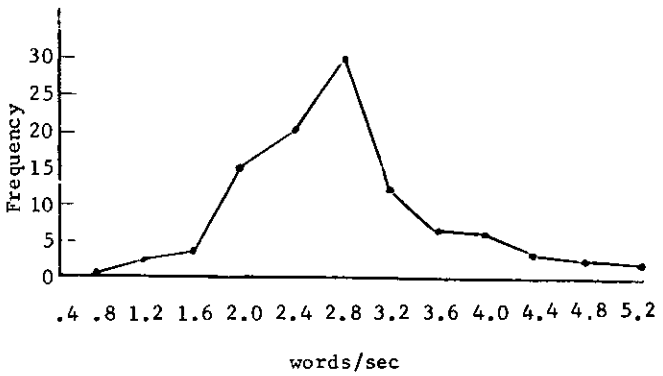


Figure A3.1: Frequency Distribution of Speech Rates.
(From Goldman-Eisler, page 19)

Goldman-Eisler measured the rates in syllables/minute and we have converted these to words/second using 1.7 syllables/word, which is only an approximate figure (Miller, 1951, p. 89). Syllable rates are probably somewhat stabler than word rates. In any case the high variability is apparent, though mean rates are in fact quite stable. Sources of variability in rates can be found everywhere. For instance, the debate that yielded the distribution above had four debaters. They were striking differences in average speech rate depending on who was talking with whom. These figures (adapted from Goldman-Eisler, page 20) provide some feeling for interpersonal variability (though on a single task). The measure is words/second.

	A	B	C	D	Rest
A		2.3	2.5	2.1	2.4
B	2.1		1.9	2.5	2.8
C	1.8	1.9		1.9	2.1
D	2.4	2.7	2.7		2.7

The column labeled Rest is conversation directed to the assembled group.

No good speech rate figures are available for tasks such as formulating communications to a machine. We do know that when people talk aloud when solving problems their speech rates drops to about 2 words/second (Newell and Simon, 1971).

Typing. The extreme figures can be taken from typing contests. The record for typing appears to be 149 words/minute, which is 2.48 words/second (N. McWhirter and R. McWhirter, 1966). The words in such a test are standardized at exactly 5 characters/word. Average typing speeds for secretaries are well known. Again, using standardized words, 60 words/minute, which is 1 word/second, is a reasonable figure. A study by Hershman and Hilliz (1965) gives some indication of the sources of variation. We do not have figures for the typing rate of engineers without typing skill using teletypes. A figure of .2 - .4 words/second would not be far off, though there is tremendous variation.

Handwriting and printing. Interestingly enough good figures could not be found in the literature for handwriting and handprinting (though norms undoubtedly exist for schoold children). Consequently, we ran a few short tests ourselves. We had five adults, secretaries and graduate students, copy two sample of technical material. We obtained the following results (in words/second):

Handwriting	.38 - .42	.39
Handprinting	.22 - .53	.35

There was little variability in the handwriting and quite a bit in handprinting. In addition there appeared to be a difference in the two samples of technical materials, one being handled uniformly at a lower rate than the other: .30 words/second for one, .44 for the other. (This averages over the two types of writing).

Telephone dialing. Extensive studies have been made of telephone dialing, both with rotary dial and push button. A recent article by Klemmer (1969) provides some basic data for push button input, which is one of the more relevant keying tasks for computer input:

Keying digits:	
Different groupings:	1.2-1.5 digits/sec
Different occupations:	.7-1.4 digits/sec

Two dimensions of variation are illustrated. The top shows that grouping by threes (365 638 591 ...) is quicker than no grouping (3 6 5 6 3 8 ...). The second line shows that clerks are appreciably faster than shop workers. (The article gives intermediate cases as well.)

Mark sense cards. Though never used in on-line operation, mark sense cards provide another useful referent point. The figure in the table (actually .43 digits/second) comes from a study by Kolesnik and Teel (1965), which compares mark sense cards with a number of other manual entry devices (a stylus punch, a thumbwheel, and a handpunch) for a particular task of entering navigational data.

A4. VOICE-DM*

DS/2 is a (keyboard I/O) data retrieval and modification system developed at SDC. The query language consists of a finite, highly constrained set of sentence frames which serve to identify a command (print, tally, etc.) and to delimit its arguments. The file for the queries is highly organized according to a hierarchy of attributes. The system responds on a CRT display. The syntax of DS/2 is given in section A4.3.

For task-DM, DS/2 has been extended to include facilities for voice input, including editing of the input stream.
These features are described in Section A4.1.
Section A4.2 has an annotated user/system protocol of the extended system.

A4.1 DS/2 Extensions for Speech Input

Symbol definitions

SYMBOL		MEANING	EXAMPLE
___	≡	speech or long pause	CHANGE__TO___ EXPLAIN____GO
-	≡	short pause	FINISH-GO
U:	≡	user input (spoken)	U: TALLY-GO
S:	≡	system output (displayed)	S: 5 entries found, request complete, Next:

DEFINE Facility

Any word not prestored in voice vocabulary must be defined at the console before it can be recognized (and therefore used in the input stream). Otherwise the user gets a "NOT DEFINED" response.

Definition process is done as follows:

DEFINE "entity" where entity is one of the following words:

DATA (BASE NAME)
REPORT
ABBREVIATION
COMPONENT
KEY
etc.

Systems asks for actual name of the entity.

User says the name.

Example:

U: SAVE REPORT ALPHA_GO
S: NOT DEFINED
U: DEFINE REPORT_GO
S: REPORT NAME IS:
U: ALPHA_GO
S: NEXT,
U: SAVE REPORT _GO | also equivalent to SAVE SAME_GO
S: NEXT,

EQUIVALENCE Facility

This allows different sounds to be recognized as the same item, value, string, or entity.

E.G. U: TALLY ENTRIES WHERE CITY EQUALS SM-PERIOD-GO
S: NOT DEFINED
U: EQUIVALENCE SM PERIOD TO SANTA MONICA-GO
S: NEXT,
U: TALLY SAME_GO
S: 5 entries found...

E.G. EQUIVALENCE BS PERIOD TO BACHELOR OF SCIENCE_GO
EQUIVALENCE IS TO EQUALS_GO
EQUIVALENCE HOW MANY TO TALLY_GO

DESCRIBE Facility

This is expanded to yield the vocal vocabulary of names, equivalence, abbreviations, etc.

E.G. U: DESCRIBE ABBREVIATIONS
S: Prints a list of abbreviations

Carriage Return - "GO"

All speech lines (strings) are terminated by a "GO" (unless they are to be deleted in which case the terminal word is "KILL").

E.G. a) U: PRINT SALARY WHERE EMPLOYEE EQUALS JONES-GO
b) U: FINISH-GO
c) U: YES-GO

Line Kill - "KILL"

A line occurs in any voiced input stream as soon as the word "KILL" is recognized. Also, KILL can perform the line kill function on a previously spoken, but not acceptable voiced input stream.

E.G. a) U: PRINT CAT WHERE COLOR EQUALS BLACK-KILL
line a) is cancelled as input

b) U: PRINT ALPHA WHERE RANGE EQUALS 1 TO 2-GO
S: ALPHA NOT DEFINED
U: KILL
line b) is cancelled as input

---

* This appendix summarized from a report by Carl Kalinowski.

Voice Input Stream Editing "REPLACE___BY___-GO"

Editing can be done on a voice input stream before that stream has been accepted (i.e., recognized and sent to DS/2 for execution). Therefore, editing a voice input stream does not change the data base as CHANGE_TO_command does.

E.G. a) U: TALLY DOG WHERE SEX EQUALS MALE-GO
S: NOT DEFINED
U: REPLACE DOG BY CAT-GO
(cat is not in voice vocabulary)
S: 12 columns required, continue (Y/N/F/B)

b) U: REPLACE TALLY BY PRINT-GO

c) U: REPLACE AND SEX EQUALS MALE BY BLANK-GO

For ease of implementation REPLACE must occur at the beginning of a voice input stream.

Punctuation

COMMA The verbal "COMMA" is used in the same places as the typewritten comma. Pauses in speech will not be interpreted as a comma -- this notion of speech segmentation must be made explicit by a voiced "COMMA". Exception: The "COMMA" is not used in numbers.

PERIOD The period will be used only to indicate that the previous string is alphanumeric or alphbetic.

E.G. U: a) PRINT VS301X5 PERIOD-GO
U: b) PRINT ABC PERIOD-GO
U: c) EQUIVALENCE BS PERIOD TO BACHELOR OF SCIENCE-GO

POINT Used with numbers to indicate decimal point.

E.G. U: TALLY WHERE DIVISIONS EQUALS ONE POINT FIVE-GO

Numbers

Numbers will be spoken according to military radiotelephony conversations (i.e., only the digits 0, 1, 2, 3 ... 9 are recognized).

E.G. a) U: PRINT BLAH WHERE BLAHA EQUALS ONE HUNDRED-GO
S: NOT DEFINED
U: REPLACE ONE HUNDRED BY ONE ZERO ZERO-GO
S: X columns required, continue (Y/N/F/B)

English	Military
ten	one zero
twenty	two zero
one thousand	one zero three
thirty-two	two point six
point six	

Relations

The following spoken words are added to the system to express the arithmetic relations.

Spoken WORD	Arithmetic Symbol
PLUS	+
MINUS	-
TIMES	x
DIVIDED BY	÷
THE QUANTITY_	(_)
QUOTE	"

E.G., FIVE PLUS THE QUANTITY SALARY TIMES TWELVE DIVIDED BY DEGREE YEAR

5 + (Salary x 12) / Degree

Operator precedence will be evaluated using standard FORTRAN conventions.

BLANK

A null is verbally expressed by a voiced word "BLANK". This is useful in deleting points of a line with REPLACE command.

E.G. U: REPLACE ZERO BY BLANK-GO

If "one zero" was spoken in the previous line, then the effect would be to change the number from 10 to 1; 100 would become 10.

A4.2 Protocol of Voice-DM

	MACHINE (Display Output)		MAN (Voice Input)
1.		1.	
2.	Enter data base information, name and volume serial number.	2.	PERSONNEL COMMA V50034 PERIOD_GO
3.	What is your security key?	3.	DEMO_GO
4.	Next;	4.	PRINT EMPLOYEES WHERE SEX IS MALE_GO
5.	Undefined print.	5.	EQUIVALENCE EMPLOYEES TO EMPLOYEE_GO

(The plural of employee was not defined in the voice recognition vocabulary; therefore, the PRINT "object" was unrecognized and required definition which was effected by the verbal EQUIVALENCE command.)

6.	Equivalence employees to employee, Next:	6. PRINT SAME_GO
7.	[Print-out] Next:	7. PRINT MAJOR WHERE HIGHDEGREE IS PHD PERIOD_GO
8.	No entries found, Next:	8. REPLACE PHD PERIOD BY MS_GO

(PHD was defined in voice vocabulary, but no such entries exist in data file; REPLACE command is used to edit verbal input string.)

9.	Undefined Replace, Next:	9. REPLACE MS BY BS PERIOD_GO

(All alphanumeric strings must be indicated as such by a succeeding verbal period.)

10.	[Print-out] Next:	10. DEFINE REPORT MS PERIOD_GO
11.	Report MS defined, Next:	11. SAVE REPORT MS PERIOD_GO

(Statement 10 defines MS to the voice recognizer as a report name and Statement 11 saves the report under normal DS operation.)

12.	Report MS saved, Next:	12. GIMME XYZ RIGHT NOW_GO
13.	Undefined request, Next:	13. KILL TALLY EMPLOYEES_GO

(KILL clears the voice input string buffer and reinitializes for input processing.)

14.	19 entries found, request complete, Next:	14. PRINT SALARY WHERE EXPERIENCE IS GREATER THAN TWENTY MONTHS_GO
15.	Undefined "greater than," Next:	15. REPLACE TWENTY BY TWO ZERO_GO

(Simplified number convention recognizes military radiotelephony numbers only.)

16.	[Print-out] Next:	16. TALLY EMPLOYEES WHERE PROJECT NUMBER TIMES THE QUANTITY ONE ZERO PLUS FIVE IS LESS THAN FIVE ZERO_GO
17.	3 entries found, request complete, Next:	17. TALLY EMPLOYEES WHERE JOB TITLE CONTAINS C-O-M-P PERIOD_GO
18.	5 entries qualified, request complete, Next:	18. PRINT SALARY COMMA JOB TITLE COMMA JOB NUMBER WHERE SAME_GO
19.	Error near field/ JOBCODE, Next:	19. EXPLAIN_GO
20.	The component is not defined for this data base, Next:	20. REPLACE JOB NUMBER BY JOBCODE_GO

(The voice recognizer contained "JOB" and "NUMBER" in its vocabulary, but the combination was inappropriate for the particular data base.)

21.	[Print-out] Next:	21. DESCRIBE HIRYR PERIOD_GO

(Interrogate data format of component HIRYR.)

22.	HIRYR FG Next:	22. HEADING F6 QUOTE HIRE YEAR QUOTE_GO

(Attempt is made to change a heading in the data base from an abbreviated form to a non-abbreviated form.)

23.	Heading F6, higher year. Next:	23. EQUIVALENCE HIRE YEAR TO H-I-R-E PERIOD BLANK Y-E-A-R PERIOD_GO
24.	Equivalence higher year to hire year, Next:	24. HEADING F6 QUOTE HIRE YEAR QUOTE_GO
25.	Heading F6, hire year, Next:	25. NEXT-KILL-LOGOUT_GO
26.	Audio logged out	

(Phonetic ambiguity arose when the voice recognizer found the pronounced "hire year" to be a concatenation of vocabulary words higher and year; to change component F6 as desired, user had to make an abbreviation equivalence where the abbreviation was actually the desired phrase spelled out in its entirety.)

A4.3 Syntax for (Written) DS/2

A formal description of the DS/2 query language is given below. This description assumes a basic understanding of formalized language notation. The notation used in this case is Backus Normal Form (BNF). A description of the symbols follows.

{	}	Choose one from the list
[	]	Optional input
[	]*	Optional, can be repeated zero or more times
<	>	Term to be defined
|		Separates alternate choices

RETRIEVAL REQUESTS

{PRINT / SUMMARY / PRINTERR / LISTSTAT} <print clause>[AND CHANGE<change clause>] [WHERE { SAME {AND / OR} <qualify clause> / <qualify clause> }]

{TALLY / CHANGE<change clause>} [WHERE { SAME {AND / OR} <qualify clause> / <qualify clause> }]

{SUBSET / PRINTERR} WHERE { SAME {AND / OR} <qualify clause> / <qualify clause> }

<print clause> ::= { ALL / SAME / <print item> } [, <print item>]*

<change clause> ::= <change item> [, <change item>]*

<qualify clause> ::= { <condition> [{AND / OR} <qualify clause>]* / <qualify clause> }

<print item> ::= { C1 / ENTRY / [<stat list>] <item> }

<item> ::= { [<id>=] { <data base component> / <expression> } / <data base component> THRU <data base component> }

<change item> ::= <data base component> TO { <value> / <data base component> / <expression> }

<stat list> ::= (AVE/AVG/COUNT/LIST/MAX/MIN/RANGE/SUM)*

<condition> ::= { <data base component> / <expression> / ENTRY } <relation> { <value> / <data base component> / <expression> }

::= { <data base component> / <expression> / ENTRY } EQ <value list>

::= { <data base component> / <expression> / ENTRY } { EQ / NQ } <value> THRU <value>

::= <data base component> CONTAINS <value list>

::= <data base component> <relation> <partial value>

<value list> ::= <value> [,<value>]*

<id> ::= a literal string no greater than 29 characters in length (enclosed in apostrophes if special characters are included).

<expression> ::= arithmetic expression using the four arithmeticoperators, data base components, and numeric constants with parentheses as required.

<data base component> ::= the name or C-number of a defined field in the data base.

<value> ::= a literal string compatible in mode to data base components in qualification.

<relation> ::= EQ/NQ/NE/LS/LT/LQ/LE/GR/GT/GQ/GE

<partial value> ::= a literal string of characters with preceding and/or trailing dots (which indicate unchecked character positions). The total number of characters and dots must equal the defined length of the component.

A5. VOICE-KP - DESCRIPTION AND EXPERIMENT*

A simulation of the Voice-KP (keypunch) system was run at RAND, using their interactive graphics programming facility.

The following experiment was run: Several subjects read a table of alphanumeric information into a microphone, with visual feedback of the results, at three different rates of random recognition error.

A5.1 System Design

The subject speaks into a microphone which is connected to a tape recorder and a remote speaker. Another person, the system monitor, listens to that remote speaker and controls the visual feedback. The subject obtains visual feedback on a CRT which echoes part of the monitor's display. Figure A5.1 shows the experimental apparatus.

The subject and monitor are placed in separate rooms so that the sound of the monitor's keyboard strokes cannot condition the response rate of the subject.

The data consists of 10 lines, each line having 10 fields of information (Figure A5.2). The data is placed on the lower half of the CRT to minimize the eye movement from sheet to display. The subject is told that we are interested in how fast the data can be entered correctly under different error conditions. Before starting the experiment, the subject becomes familiar with the system in a test run with 3 lines of different data.

The subject's display contains the 10 column headings; after each field value is spoken, the response value appears in the appropriate column. For each response, one of four possible errors might occur:

E1: an incorrect value is given; (the system chooses the incorrect value randomly from a list of 3 incorrect values stored with each correct value). The visual feedback for this condition is just the incorrect value itself. This condition corresponds to the system thinking it could recognize the input speech, but performing incorrect recognition.

E2: the system knows it doesn't recognize the input speech. The visual feedback is a column entry of "?????", with asterisks under the field. Asterisks appearing under any field indicate that the system knows that column entry is in error.

E3: an incorrect value is given, and the system is so confused that it will not give a correct value again (for any column) until it is "reset". There is no visual feedback distinguishing this condition from condition E1, except that consistently incorrect responses are received.

E4: the system knows it doesn't recognize the input speech, and knows it is so confused that it will not give a correct value again (for any column) until it is "reset." The visual feedback is asterisks appearing under all remaining fields on the line.

The subject's speech is restricted to the following format: he can mention data values, in order, from left-to-right; to "reposition" himself at any column, he can mention the column heading. For example, given the column headings A, B, C, D, E and corresponding data values a,b,c,d,e, the following two dialogs are correct (feedback values for each response are given in parentheses).

first dialog:

a (a) b (x) B b (b) c (c) d (d) e (e);

second dialog:

a (a) b (x) c (c) B b (b) D d (d) e (e);

In addition to column headings and data values, the subject has two commands available:

RESET - to reset the system if he thinks or knows the system is "hopelessly confused";

NEXT - when he thinks the displayed values for an entire line are correct, this command erases that line, and the system awaits values for the next line.

Each subject made 3 runs, each with different error probabilities. The system flow is indicated in Figure A5.3. (Let p(Ei) be the probability that condition Ei will occur.) Therefore, if the system is not already hopelessly confused, the probability that some

---

* This experiment and report was done by Robert H. Anderson.

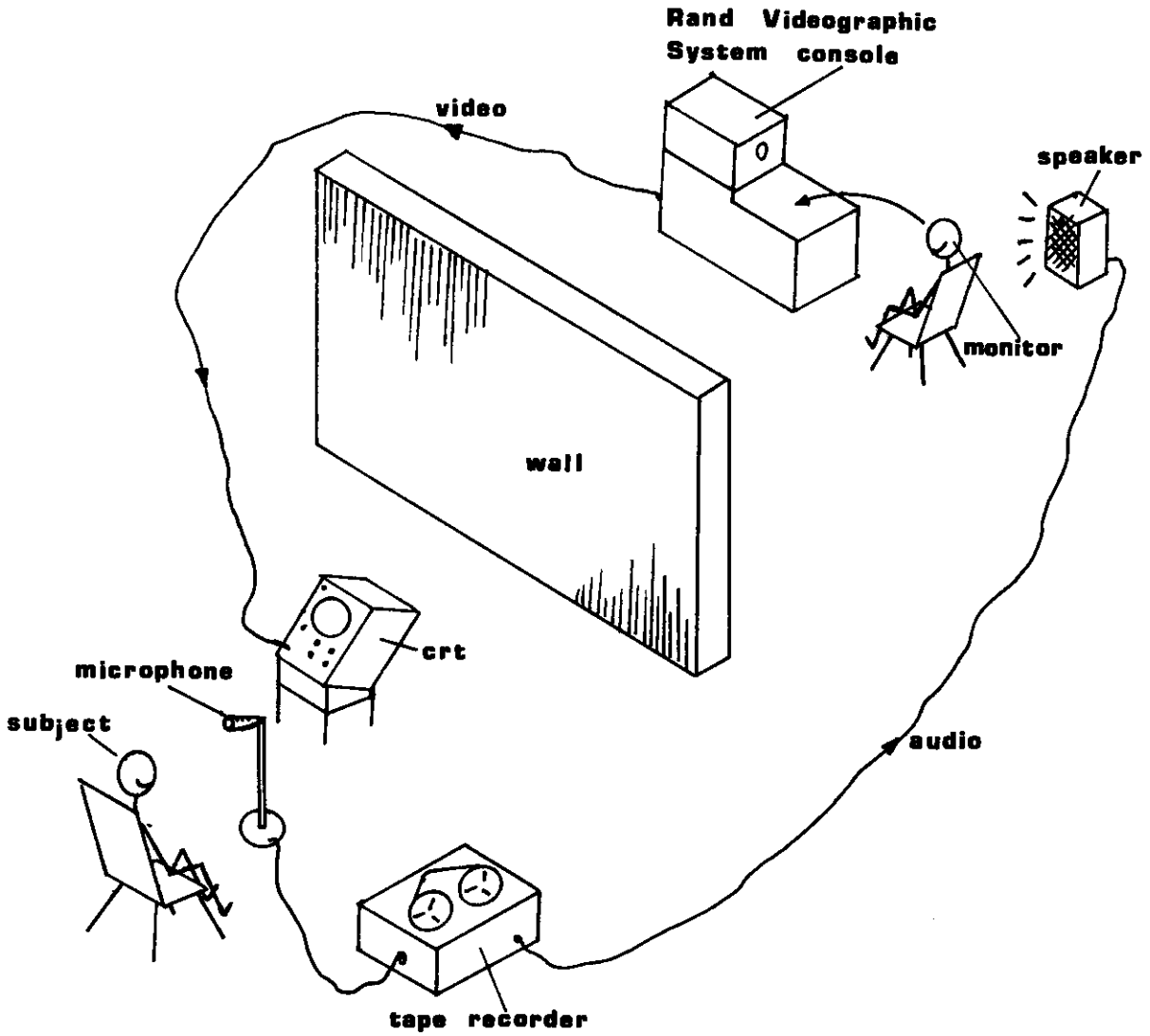


Figure A5.1 EXPERIMENTAL APPARATUS

EMPLNO	SURNAME	INITIAL	SEX	AGE	MARSTAT	DEP	DRAFT	DEG	MAJOR
00365	CHARLSTON	G	M	37	M	0	3A	MA	MATH
00366	MC-GAHEY	N	M	31	M	2	3A	BA	MATH
00377	SMART	J	M	36	S	0	1A	BA	ENGINEER
00398	WEBSTER-II	D	M	38	M	2	5A	BA	POLSCIENCE
00469	CASSLEY	W	M	42	M	1	3A	BA	ENGLISH
00470	LONG	J	M	43	M	2	3A	MA	BUSADMIN
00561	CLARK	L	M	35	M	0	5A	BA	POLSCIENCE
00572	BLACK	R	M	37	M	2	2A	PHD	MATH
03284	CALLAHAN	R	M	34	M	2	5A	MA	MUSIC
05289	COCHRAN-JR	C	M	32	S	0	1Y	BA	MATH

Figure A5.2: Data for input by subject.

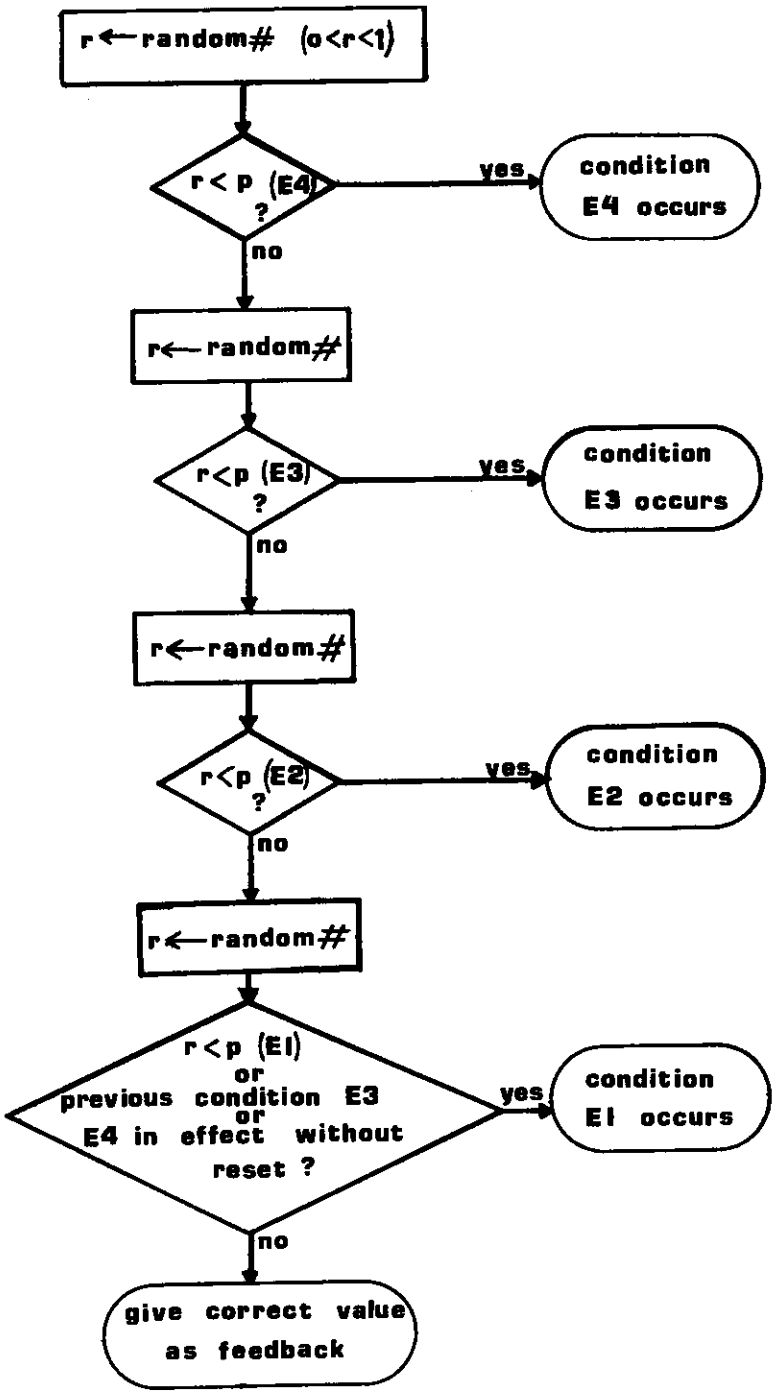


Figure A5.3 Flowchart of Simulation

error will occur in a particular response is

$$P(E1) + p(E2) + p(E3) + p(E4)$$

The error probabilities for the 3 runs were:

Run #	p(E1)	p(E2)	p(E3)	p(E4)	Σp(Ei)
1	10%	10%	2%	2%	24%
2	5%	5%	1%	1%	12%
3	2.5%	2.5%	0.5%	0.5%	6%

We felt that a 24% error rate was near the upper tolerance level for a recognition system; the slow data rate at this error rate bears out this feeling; the 6% error rate permitted nearly error-free input with little interruption of the data flow.

A5.2 Results of the Experiment

The elapsed time for each run was obtained from the tape recording of the session using a stopwatch. The elapsed times (in minutes) are:

	Run #1	Run #2	Run #3
Subject #1	7.2	5.1	4.9
Subject #2	6.1	5.1	4.1
Subject #3	5.0	5.1	4.2

In order to place these elapsed times in the proper perspective, some additional measurements were made using the same 10 lines of input data:

1) subject #1, with the standard visual feedback, but 0% error rate: 2.8 minutes;

2) Dr. Raj Reddy:

   24% error rate: 9.0 minutes,
   12% error rate: 5.8 minutes;
   0% error rate, without visual feedback (i.e., reciting data values into an assumed perfect recognizer, with no verification): 1.8 minutes

3) Data was submitted to Rand's keypunching service. The elapsed times were:

   keypunch 10 cards: 3 1/2 minutes
   repunch 10 cards : 2 minutes
   reconcile both decks (using an IBM 519 Reproducer to compare them) until they agree : 3 1/2 minutes
   9 minutes

   (These times are quite informal. The operators timed themselves. The 3.5 minute punch time probably includes punching the format card.)

4) Data was submitted to a secretary who is an excellent typist. Elapsed time to type the information, with no errors:

   typing: 4 minutes
   proofreading: 0.5 minutes
   4.5 minutes

All of the above timing information is summarized in Figure A5.4.

As each run was being made, a historical file was created showing all feedbacks the subject saw. A record was written into this file each time the monitor responded to the voice input.

A5.3 Discussion

There is not enough data, and the data is not sufficiently clustered, to form firm conclusions about recognition of continuous speech. In analyzing this data, the following factors should be considered:

1) The discontinuity of the data. It takes more time to spot errors in 5-digit numbers than in, for example, English text. Also, after looking at the visual feedback, it is difficult to find one's place again in a data table with little context.

2) System response time. The average feedback delay for each column entry was about one second, with a range from about 0.5 seconds to about 1.5 seconds. The subject tended to pace himself on this feedback, which tended to slow down the data rate. An indication of this effect is that Raj Reddy took 1.8 minutes to read the data with no feedback at all, and subject #1 took 2.8 minutes to read the data with feedback but 0% error rate. The one minute difference can be attributed to a hesitation of about 0.6 seconds per data time while awaiting the feedback response.

3) Visual feedback. Reading unfamiliar data from a table and verifying it from visual feedback is a "worst case" situation in which the subject's eyes are continuously moving from data to feedback and back. This experiment has little bearing on the use of voice recognition in situations where the input data is not being read, but is "self-generated" (e.g., a programmer dictating a program over the telephone, a fighter pilot stating range estimations for his programmed missiles). Another aspect of the visual feedback was its resolution. Seventy-two characters of information were displayed in 8 inches on a raster-scan

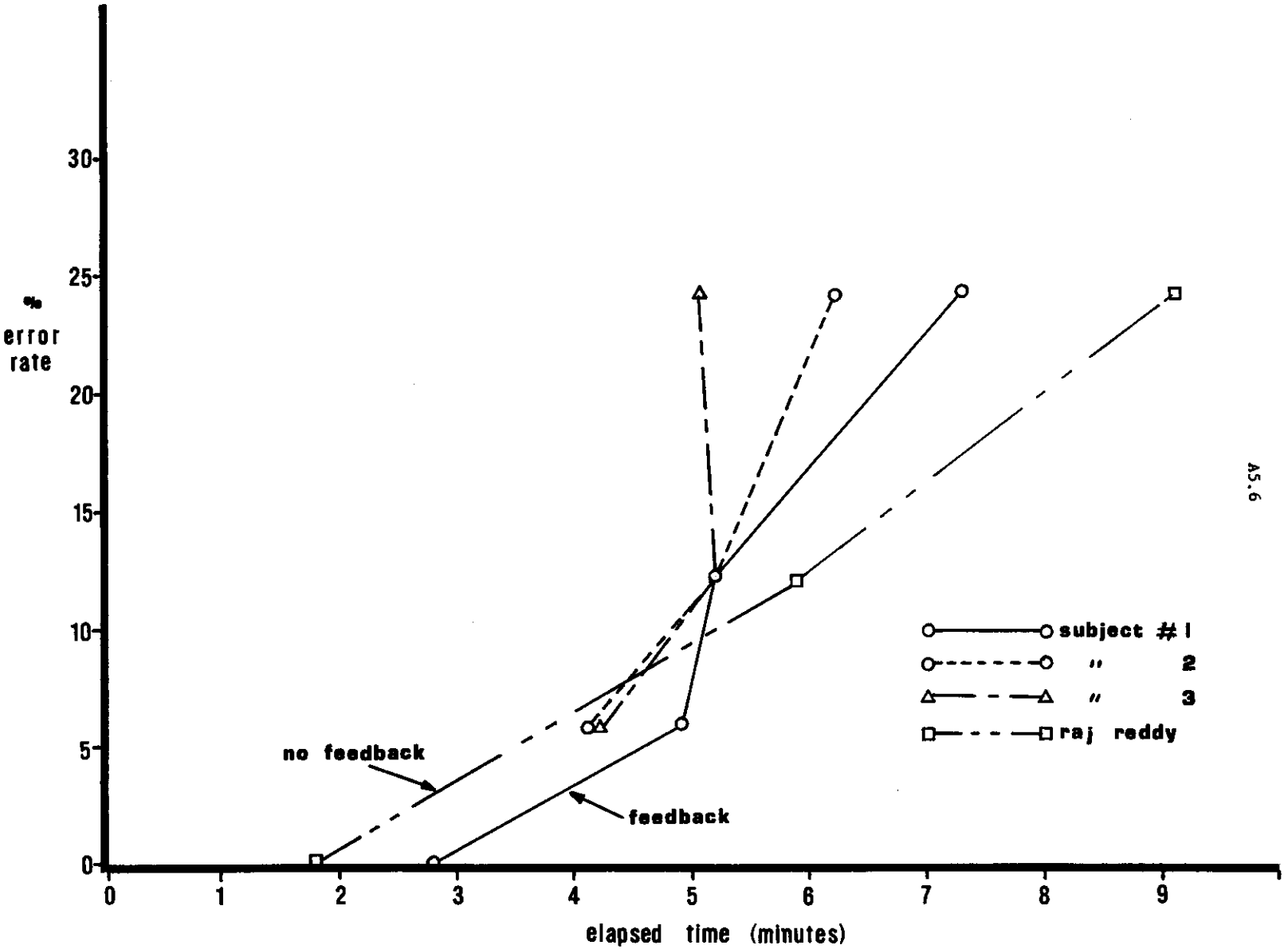


Figure A5.4 ELAPSED TIME FOR INPUT vs. ERROR RATE CONDITION.

refreshed CRT. The similarity between "M" and "N" or "D" and "B" under these conditions requires rather detailed scrutiny of the feedback. In a more careful experiment, the feedback should be displayed on several lines, or possibly vertically, to allow a larger character size to be used.

A6. VOICE-CS*

This appendix describes in detail the following aspects of the Voice-CS project:

1. the data base;
2. the vocabulary;
3. the syntax; and
4. a protocol.

A6.1 The Data Base

Following are the types of information about the system that Voice-CS provides. Figure A6.1 illustrates the types of information provided by the current SYSTAT program.

1. Statistics relating directly to users' jobs. For each job on the system, the following may be given: job number; user's identification; where the job is logged in; the name of the program being run; the size of the program; the run status of the job; and the accumulated run-time of the job.

2. What devices are available.

3. General system statistics, including uptime, nulltime, virtual core used, and real core used.

A6.2 Input Vocabulary

Here is the set of keywords that Voice-CS recognizes. These words convey the basic meanings of input utterances. In actual operation, Voice-CS needs to rely additionally on a set of secondary words, such as "and" and "the" in order to confirm recognition done on the basis of this primary vocabulary.

ARE	I/O-WAIT
AVAILABLE	IS
BUSY	JOB
COMPUTER	JOBS
CONTROL-C	K
CORE	MAGNETIC
DATE	MAGTAPE
DECTAPE	MEMORY
DISK	MINUTES
DOWN	MTA
DTA	NINE
EIGHT	NOT
FIVE	NULLTIME
FOUR	NUMBER
FREE	ON
HOURS	ONE
HOW-LONG	PDP10
HOW-MANY	PEOPLE
HOW-MUCH	PRINTER
PROGRAM	TELETYPE
PROGRAMMER	THREE
PROJECT	TIME
RESOURCES	TTY
RUNTIME	T.T.Y.
SECONDS	TWO
SEVEN	UP
SIX	UPTIME
SIZE	USERS
SPACE	VIRTUAL
STATE	WHAT
STATUS	WHERE
SWAPPING	WHO
SYSTEM	WORKING
TAPE	ZERO

A6.3 Input Syntax

Below is a simplified input syntax for Voice-CS. A number of further constructs need to be added to correspond to possible synonyms to be added to the vocabulary.

Unlike ordinary BNF, these productions do not necessarily imply the order of the symbols given in them. Thus the production

<reference> ::= <noun> <n-qualifier>

implies the additional (but perhaps less likely) production

<reference> ::= <n-qualifier> <noun>

In addition, words which happen to be adjacent in the grammar need not be adjacent in actual input.

```
<question> ::= <yes-no>
               <information-q>

<yes-no> ::= <be> <reference>
             <be> <reference> <availability>

<be> ::= IS
         ARE

<information-q> ::= <q-phrase> <reference>
                    <q-phrase> <reference>
                               <availability>

<q-phrase> ::= <q-word>
               <q-word> <q-qualifier>

<q-word> ::= WHAT
             WHO
             WHERE
             HOW-MUCH
             HOW-MANY
             HOW-LONG
```

* This appendix was produced by R. Neely and L. Erman.

```
JOB 8       CMU.4   -   DEC  4S72.F

STATUS OF CMU.4     -   DEC 4S72.F AT 10:59:05 ON 03-JUN-70

UPTIME 07:24:05, 96% NULL TIME = IDLE+LOST = 96% + 0%

JOB     WHO         WHERE    WHAT     SIZE     STATE     RUNTIME

1       N110CG15    TTY0     XBLISS   12K      TT        00:08:49

2       A610LE03    TTY7     LST360   12K      ↑C SW     00:00:35

3       A610RF07    TTY5     LINED    4K       TT        00:00:40

4       E210FN02    TTY1     LINED    4K       TT        00:00:14

5       3,3         DET      PRNTR3   5K       SL        00:00:06

6       A110DM30    TTY2     LSD29    15K      TT SW     00:02:21

7       N605JT29    TTY6     PIP      4K       TT        00:02:12

8       2,4         CTY      SYSTAT   2K       RN        00:00:00

HIGH SEGMENTS

PROGRAM OWNER       HIGH K    USERS

LINED   1,1         6K        2
(PRIV)  JOB 1       20K       1

DORMANT SEGMENTS

PROGRAM OWNER       HIGH K

LOGIN     1,1       1K SW
COMPIL    1,1       2K
LOGOUT    1,1       1K SW
BASIC     1,1       5K SW
LOADER    1,1       2K SW
MACRO     1,1       5K SW
TECO      1,1       2K

% SWAPPING SPACE USED = 51/200 = 26%
% VIRT. CORE USED = 84/200 = 42%
7K CORE LEFT

%VIRT. CORE SAVED BY SHARING = 6/(6+84) = 7%

BUSY DEVICES:

DEVICE      JOB       WHY

LPT         3         INIT
DTA1        4         AS
DTA2        4         AS
DTA6        6         AS
DTA7        7         AS
MTA0        1         AS
```

Figure A6.1. Information provided by the current SYSTAT program

```
<q-qualifier> ::= JOB NUMBER
                  PROJECT PROGRAMMER NUMBER
                  PROGRAM
                  STATUS
                  STATE
                  RUNTIME
                  <terminal>
                  <bigness>

<reference> ::= <noun>
                <noun> <n-qualifier>
                <n-qualifier>

<availability> ::= FREE
                   AVAILABLE
                   BUSY

<terminal> ::= <terminal-dev>
               ON <terminal-dev>

<terminal-dev> ::= TELETYPE
                   TTY
                   T.T.Y.

<bigness> ::= <real-bigness>
              VIRTUAL <real-bigness>

<real-bigness> ::= SIZE
                   CORE
                   SPACE
                   MEMORY

<noun> ::= JOBS
           PEOPLE
           USERS
           COMPUTER
           SYSTEM
           PDP10
           NULLTIME
           UPTIME
           DATE
           TIME
           RESOURCES
           JOB NUMBER
           PROJECT PROGRAMMER NUMBER
           STATUS
           STATE
           RUNTIME
           DISK
           DECTAPE
           DTA
           MAGTAPE
           MAGNETIC TAPE
           MTA
           PRINTER
           <terminal>
<n-qualifier> ::= <adjective>
                  NOT <adjective>

<adjective> ::= <non-neg integer>
                <non-neg integer> K
                <non-neg integer> HOURS
                   <non-neg integer> MINUTES
                   <non-neg integer> SECONDS
                CONTROL-C
                I/O-WAIT
                DOWN
                ON
                SWAPPING
                UP
                WORKING

<non-neg integer> ::= <digit>
                      <digit> <non-neg integer>

<digit> ::= ZERO
            ONE
            TWO
            THREE
            FOUR
            FIVE
            SIX
            SEVEN
            EIGHT
            NINE
```

A6.4 Protocol

The protocols given here and in Chapter 6 were obtained in an attempted simulation of an ideal Voice-CS program. For this simulation a tape recording was made of several short conversations in which a small set of PDP-10 users asked questions of a human experimenter. The purpose of the protocols is to get an idea of the types of questions to expect from a user. The experimenter's answers in fact are often not confined to what could be expected from Voice-CS.

U: The system up and running?

M: Hello this is the PDP-10.

U: Oh, is the system running now?

M: The system is running

U: Ah.. are there any ah DEC tapes available?

M: Available DEC tapes are DTA1 DTA2 DTA3 DTA7.

U: How much disc space is there? (bb)

M: There are 57 blocks left.

U: (b) Thank you.

M: Goodbye.

U: Bye.

* * * * * * * * *

M: Hello this is the PDP-10.

U: Ha.ha.. Ok.. um.(b).. let see.. wh.what things would I want to know right off hand -- I'd like to know how many users are on the line? On.. are on the system that's obvious.

M: Could you please rephrase that question?

U: (b)... How many users ah are there?

M: There are now 37 users on the system.

U: OK. mmm..mm... How many backtra. how many background jobs are there?

M: Please rephrase that question I did not understand?

U: Oh dear... (b) how many detached jobs are there?

M: Five jobs are detached.

U: OK... ah What is the average response time in the system?

M: I do not understand that question.

U: OK... ah I guess.. what um percentage of the CPU. um.. could I get if I had a full compute bound job?

M: Currently 37 percent null times.

U: ...(b) It's not quite the answer I wanted... but... (b)

M: I do not understand that. Would you like me to call my supervisor?

U: (b)... aaahhhm.....

M: Hello are you still there?

U: Yeah...I'm still there.. ah.. (b) ah... what else would I like to know (very low) ... mmm are there any DEC tape drives available? Are there two DEC trape (sic) drives available? mmm..

M: I do not understand the question, please rephrase?

U: (b) mm.mm Are there two..free.. DEC tape drives?

M: DEC tapes 1, 3, 7 and zero are free.

U: (low) OK..... OK Off the top of my head I can't think of any more, but

M: I do not understand that.

U: (Laughter) (b) tough... umm... OK, that is all I want to know... Period. And hanging up will cut off this conversation.

M: I do not understand, please rephrase.

U: Nope.

* * * * * * * * *

M: Hello this is the PDP-10.

U: How's the computer feeling today?

M: I don't understand that, should I get my supervisor?

U: Let me rephrase that question, um.. can I come to work today?

M: I don't understand that.

U: Ah ... ah ah how many jobs are logged in?

M: 433 jobs are currently logged in.

U: er ah... I assume the system is up.

M: The system is up.

U: Amazing... um.. OK are there any free teletypes?

M: Teletypes 3, 4, 5, 6, 7 and 8 are free.

U: OK, are there any display lines?

M: Teletypes 3, 4, 5, 6, 7 and 8 are free.

U: Er... OK, thank you Mr. PDP-10.

M: Are you saying goodbye?

U: Right...

M: Goodbye.

* * * * * * * * *

U: Hello

M: Hello, this is the PDP-10.

U: Oh, I would like some... information about the... computer system please.

M: This is the PDP-10 time sharing system.

U: Ah, ah could you tell me how many people are on the system today?

M: 143 people are currently logged in.

U: My goodness, is is is it possible for me to get in?

M: I don't understand that question, shall I call my supervisor?

U: No, my, perhaps you could tell me... ah.. what the null time is on the system.

M: We are currently running 33 percent null time.

U: I see. Could you also tell me ah.. how many.. what is the.. the extend. extended ah.. no I'm sorry I didn't mean to say that. Could you tell also how much... the core exceeds by the requests?

M: We are currently running with 112K of core, there are 43K of swapping core left. The core is exceeded by 150K.

U: Thank you very much. I appreciate it. Bye bye.

M: Goodbye.

* * * * * * * * *

A7. VOICE-CC*

The computer consultant task is to provide interactive information to a user attempting to run under the new TENEX system on the PDP-10 at Bolt, Beranek and Newman. The user is assumed to be familiar with computers (and with some time-sharing system), but is a novice to this system. The user converses with Voice-CC over a voice channel as he attempts to use the TENEX system over a conventional (teletype or display) terminal; the Voice-CC responses also come over the auditory channel. In addition to the verbal input, Voice-CC receives information by monitoring the user's interaction with TENEX.

The input language is a sophisticated artificial language that is highly English-like. The user receives no training specific to this language; his use of it is naturally shaped by interaction with the Voice-CC system. Some possible restrictions on the language are given in section A7.2.

A7.1 Recorded Protocol for Voice-CC

The accompanying protocol gives an impression of the type of spoken interaction that might occur in an ideal system for voice assistance to an on-line user of a computer system--in this case, the TENEX system of Bolt, Beranek and Newman, Inc. This protocol is based on a tape recorded session between one of the designers and implementers of the TENEX time-sharing system and a user who had not previously used the TENEX system (although he was familiar with another interactive time-sharing system and was an experienced programmer). The purpose of the protocol is to get a realistic picture of the phonetic, syntactic, and semantic characteristics of the questions which a user might ask in such an environment, unconstrained at this point by considerations of the limitations of a computer question-answering system using the present technology. The ideal goal is to make the computer information system as natural to use as the human expert who "simulated" the computer in this study.

In the accompanying protocol, the questions of the user are represented as closely as possible to the way in which they occur on the tape. In addition, there are comments in the margin concerning aspects of the sounds that appear on the tape. (There are a few places on the tape where some discussion takes place between the user and the systems programmer concerning the ways in which a computer might be expected to answer the question. These aside comments have been omitted from the typed protocol as they are not the type of interaction which we are contemplating automating.) The replies of the TENEX systems programmer who simulated the computer have been edited from the form in which they actually occur on the tape (informal, definitely human communication) to correspond to the type of prose that one might expect from the computer in an ideal system (and also for brevity since it is the user's questions which primarily concern us). No attempt has been made, however, to limit these replies to the range of knowledge or the sophistication of response generation that are currently state-of-the-art in question-answering. Thus, the task of actually generating such replies by a machine is a non-trivial task (as are the tasks of recognizing the speech and understanding the question).

Protocol for Voice-CC

1. User: How do I log onto TENEX?

2. Computer: Type CONTROL C followed by LOG followed by ALT MODE

3. User: CONTROL C -- noise -- followed by LOG followed by ALT MODE -- noise -- Now what do I do?

4. Computer: Type your user name followed by ALT MODE.

5. User: Okay

6. Computer: Beyond this point I will dispense with saying "terminated by ALT MODE." Every executive command in TENEX can be terminated with either ALT MODE or carriage return. ALT MODE will cause printing of any implicit information that belongs on the line while carriage return will not.

7. User: Okay, ah, user name -- noise -- Do I have to, ah, put a period or anything?

8. Computer: You can terminate it with a carriage return.

9. User: -- noise -- Okay, what's the password? Ah, carriage return again?

   (The question "what's the password" is directed by the user to himself paraphrasing the current demand by the teletype. A switch could be used to tell the system when it's being addressed.)

---

* This appendix was produced by W. A. Woods.

10. Computer: Yes

11. User: -- noise ---------- okay, um, how can I look at my file directory?

12. Computer: Type DIRECTORY followed by carriage return.

13. User: -- noise ----- ah, okay is there some way I can kill that in midstream?

14. Computer: Type CONTROL C.

15. User: Okay, CONTROL C doesn't hurt anything when I type it?

16. Computer: CONTROL C will stop any job in any program you have running with the possibility of continuing it later, or it will stop an EXEC command and return you to the EXEC command level.

17. User: I see, how do I continue a job I've stopped with CONTROL C?

18. Computer: Type carriage return.

19. User: -- noise ------ okay, but that doesn't work with an EXEC command?

20. Computer: No, that's right, it only works with a program.

21. User: --pause--ah, fine, how can I list a file to the tele... ah, line printer?

22. Computer: Type the EXEC command LIST followed by space and the file name that you want listed.

23. User: -- noise -- do I terminate it with anything?

24. Computer: carriage return

25. User: -- noise -- what does it want now?

26. Computer: It is now listing the file. The dates appearing on your teletype are due to a temporary bug in the system. When it's finished it will respond with the @ sign in the left margin.

27. User: Ah, when the bug is fixed will it type anything at all there where the date stands or it won't say anything?

28. Computer: It says nothing until it's finished listing.

29. User: How long does it take to, ah, list a file?

30. Computer: The line printer runs at 600 lines a minute, which is about 10 pages a minute.

31. User: Do I have to worry about whether the line printer is offline or online or anything like that?

32. Computer: No, if any unusual conditions exist it will tell you.

33. User: Okay

    -- long pause --

34. User: an' un, I suppose I go and pickup the printout from the line printer when it's ... when it's through.

35. Computer: Right.

    -- pause --

36. User: It seems to be taking an awful long time.

37. Computer: What size file is that?

38. User: It's maybe -- pause -- 15 pages or something like that.

    (The statement trails off with "or something like that."

    -- long pause --

39. User: Does this list command . . . space pages?

40. Computer: No.

41. User: It runs right across the page boundary?

42. Computer: It puts page headings only at page divisions within a file; it doesn't break pages in logical places.

43. User: Uh, it there a facility that will do that?

44. Computer: Not yet, the LIST command will do that eventually.

45. User: I see. Ah, there's a list-file command in LISP - does that do it?

46. Computer: No, it doesn't at present.

47. User: (It) used to on the 940.

    (The initial "it" is almost invisible on the tape.)

48. Computer: All it actually did was call utility.

49. User: Ah! (laugh) -- noise --

50. Computer: There we are.

51. User: Is there a way to, to ask for the status of the system? (repeated word "to")

52. Computer: Not without stopping the listing that's currently going on.

53. User: I see.

    -- pause -- noise --

54. User: Okay, oh, I see, it does an echo. (almost indistinguishable from "it doesn't echo.")

55. Computer: Yes

56. User: It won't type before the @ sign when it's fixed though? (question trails off at end)

57. Computer: Normally that @ sign should appear in the left margin.

58. User: Okay, ah, what kind of things can you do with a file directory? -- You can delete files. . . . can you copy files?

59. Computer: Yes, you can copy files.

60. User: Ah, how do I create a new file that's a copy of an old one?

61. Computer: You can use the copy command which is of the form COPY <file> to <new file>.

62. User: -- noise -- What does the period mean?

63. Computer: A period normally separates the file name from the file extension - in this case your files don't have extensions so the period has nothing following it.

64. User: I see. (The "I" is drawn out for several seconds.)

65. Computer: Also at this point if you were typing a file name that exists already you would need only to type as much of it as will uniquely identify it. When you think you have enough you can type ALT MODE and the system will print out the rest of the name. You can do that whether you are reading or writing, except that in the case of a new file you have to supply the whole name and indicate with a space or carriage return that you have completed the name.

66. User: Okay.

67. Computer: At this point the system will respond with "new file" or "old file."

68. User: Um, does . . .
    If I say I think I'm through but there isn't enough to determine the name then what will the system do?

69. Computer: In that case it will ring the bell and will allow you to add some more.

70. User: I see.

71. Computer: The same command recognition also applies to command names for the exec. Why don't you start this command over again? Type CONTROL C.

72. User: -- noise --

73. Computer: Type something you think would identify the COPY command followed by an ALT MODE.

74. User: -- noise -- Okay -- noise ------ 'nother space?

75. Computer: If you typed ALT MODE for that file name, then you would have gotten more information out of the system about how the copy command is formatted.

76. User: Uh huh, okay -- noise ------- now, what does that tell me?

77. Computer: That says it's recognized the file name which you typed in - it says no extension version one.

78. User: I see. So, that's what it should normally do if I typed a space as well or . . .

79. Computer: No, when you type a space it doesn't supply the recognition printout that it does in the case of ALT MODE. It simply goes on to the next thing.

A8. OTHER POSSIBLE TASKS FOR SPEECH-UNDERSTANDING SYSTEMS

The airline-guide information service task might involve planning a complex trip or might only require a simple table look-up. The input language is free English, but much use of a small sublanguage for talking about airlines, etc., may be expected. The output language (speech) may be completely stereotyped vis a vis this sublanguage. The interaction will hopefully be real-time, although this is not a critical constraint. Although useful variants of this task may exist, the task per se is not clearly useful.

The desk calculator task does the job of (say) a standard electronic desk calculator, namely accepting commands, performing calculations, and storing results. The input language is restricted to the technical sublanguage relating to the task. The spoken output is also simple and needs to be augmented with visual output. The constraints on the interaction are similar to those of the airline-guide information service task. This task is not particularly useful, since it is not clearly better than graphic alternatives.

The air traffic controller task consists of real decisions to be made in terms of a dynamic model of the world. The input speech will be entirely in a technical sublanguage but with deviation in utterance under stress. The speech output is stylized as with the airline-guide information service task. Unlike that task, the real-time constraint is critical here because of the nature of the air traffic controller problem. Speech will be input in a noisy environment. The semantic model is non-trivial but fairly easy. This is a highly relevant task, but may be too hard. In any case, voice communication is inherent.

In the missile checkout task the man goes through a checkout procedure, speaking his observations to a computer and answering questions from the computer. The task content is simply following a large decision tree, but is good in that it leaves the computer in control of context. The speech input will be in a technical language, with the speaker especially trained for the task. Speech output may be totally stereotyped, with a specific utterance for each point in the decision tree. This is also a relevant task, and is interesting, since the roles of man and computer are reversed from the other systems.

In the medical history taking task the new patient gives a medical history. This task does not include undertaking a diagnosis of the patient, although it should request elaboration in special areas on the basis of conclusions made from its "knowledge" of medicine. Thus, the task requires minimal intelligence. The input is free speech, but mostly short comments (or it can be forced to this). The speech output is highly stylized, rather Eliza-like. This is a natural task in that people would ordinarily rather talk than write.

The automatic protocol analysis task is inherently an artificial intelligence program designed to develop hypotheses of problem-solving behavior. The speech input is free but rather simple English. There is no speech output and no real-time constraint. The semantic model is interesting -- namely, the theory of the subject involved in problem-solving. This task, of course, is of interest only to those select few concerned with this area of interest.

The physical inventorying task involves a single person moving through a warehouse taking inventory on the items. The use of voice input allows him to keep his hands free for moving and touching the physical objects. The interaction is constrained in much the same way as in the data management query task. Actually, the constraints are even stronger since the system itself can know the general layout of the warehouse and therefore know what is to be talked about next. It could even provide master control, determining which bin was to be considered next and promptiag the human.

The robot management task involves giving a robot verbal instructions about how to move and behave in a cooperative task in, say, a room. Here again, the physical freedom of movement of the human and of the robot makes the use of voice attractive. The fact that the human and the robot are both focussed on the same task, for which the robot has independent information, provides some additional semantic support that might be exploitable.

A9. ANALYSIS OF THE TASKS

Chapter 6 summarizes the results obtained through various analyses and comparisons of the tasks. This appendix provides more detailed descriptions of the analyses to substantiate statements made in that chapter.

The organization of this appendix follows closely that of Chapter 6 and will contain mostly supplementary material. However, some parts may be repeated in the appendix to make it self-contained by itself.

A9.1 System Organization

Most speech understanding systems are organized into levels corresponding roughly to the levels recognized in linguistics and acoustic research. Here we label these levels as semantic, sentence, lexical, phonemic, parametric and acoustic levels. We will illustrate the nature of these levels by considering specific examples from the Voice-CS system. Figure A9.1 provides a summary of the levels, their representations and the sources of knowledge.

Semantic Level: There are three independent semantic structures in Voice-CS. The first is the status information on the PDP-10. This is kept by the PDP-10 Monitor and Voice-CS has no special responsibilities for it. It has access to it via a set of commands to the monitor for specific items of data, whose details need not concern us. Since its structure is fixed, we can assume that whatever processing is appropriate is simply built into Voice-CS.

The second semantic structure is a representation of the user's desires for status information. We adopt a simple view of the requests that the user can make. Our representation for this, called the elementary sentence form, is in fact the representation at the sentence level (which we discuss below). Thus, there is no separate representation at the semantic level for the user's requests for information.

The third component of the semantic representation is a model of the user's communication state, which is a necessary part of any conversational system. Without this, no element of grace or consideration can enter into the conversation. Ultimately, one might expect to develop a psychological model of the user, from which his responses could be predicted and also his reactions to the system's statements. Essentially nothing has been done in characterizing conversations in a way useful for man-computer conversations (for the kind of thing that has been done, see Goffman, 1967).

A standard device, useful in such situations of ignorance, is to create a finite collection of states, each standing for a conventional "position" that the user can be in vis a vis Voice-CS. Figure A9.2 gives a moderately appropriate example. At any moment Voice-CS takes the user as being in one of these states, e.g., as just having initiated the conversations, or as having become confused, or as a new user who does not know what can be asked of Voice-CS.

The usefulness of such a state system lies in whether different actions are appropriate to a user in different states or (more important for use) whether different utterances can be expected in different states, thus establishing a limited context for recognition. For example, in the state, repeating, the just prior utterance may provide a good guide to the present utterance. The transitions from one state to another are either derived from the logic of the situation or from past statistics of transitions.

Sentence Level: The second level in Figure A9.1 is the sentence level. The simplicity of task permits us to force all requests into a simple Procrustean bed, indicated by the simple schema:

(COUNT)ATTRIBUTE(OBJECT) = VALUE

The status tables define a set of objects: the system, individual jobs, resources such as tapes and printers, users, programs, and the report itself. These objects have various attributes, which can take on various values. It is usually these values that the user is requesting. Figure A9.3 shows a number of examples, giving the answer to the request at the far right. Thus, an attribute of the system is its uptime (item 1), and one can request its value, which is the number of hours that the system has been running. Another attribute (item 2) is the DECtapes on the system, and one can request this value, which is a particular set. One could also request a count of this set (item 3), or even a yes/no answer to whether there were two DECtapes available (item 4).

All possible requests for status information can be expressed by filling in (or leaving blank) the four items. This is the elementary sentence form, and it plays the same role for Voice-CS that a parsed sentence does for a more complex task using a more elaborate language. Thus, the sentence level consists of a sequence of instances of such elementary sentence forms. Note that this is a design decision. The user, in fact, may have other desires for status information which Voice-CS does not recognize.

An important specification for Voice-CS is that the user may use unconstrained English over the telephone. Many of the words spoken lie outside the vocabulary used in the elementary sentence form. Voice-CS uses a modified form of

Semantic level

System's status

Represented by table in PDP10, accessible via Monitor
Fixed structure, known at design time, built into Voice-CS

User's desires for status information

Represented by elementary sentence form, hence no specific semantic representation.
Elementary form fixed by design, built into Voice-CS
Frequency of requests determined by experience

User's communication state

Represented by finite state system
Fixed state system determined by logic of conversation
Frequency of transitions determined by experience

Sentence level

Represented by elementary form: (COUNT) ATTRIBUTE(OBJECT)=VALUE
Role of each word (syntax-semantics dictionary) for form determined by knowledge of English grammar and semantics.
Simple word order rules of English
Frequency of word orders determined by experience

Lexical level

Represented by sequences of words
Finite set of words in dictionary with one (possibly more) phonemic sequence for each.
From standard knowledge of English phonetics
Phonological rules (including conversational transformations)
Stress and intonation rules
Phoneme order statistics
A priori from English
Calculated for local languages

Phonemic level

Represented by sequences of phoneme-lists, where each phoneme list gives the alternative phonemes that could occur at a given point, ordered in likelihood of occurrence.
Parametric representations for each phoneme
Base parametric representations for each phoneme
Co-articulatory rules
General rules of continuity for phonemes

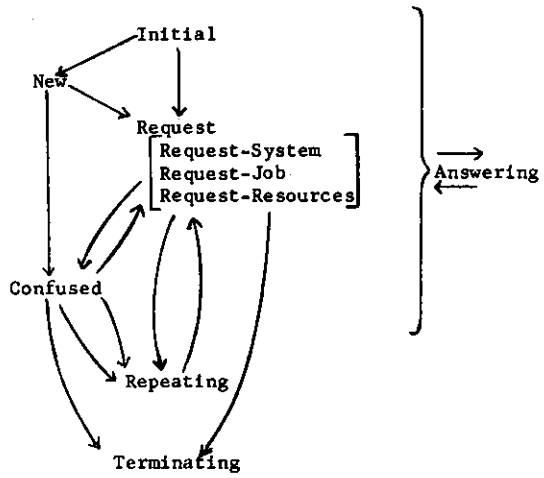
Parametric level

Represented by sequence of parallel measurements
Articulatory rules for significant parameters of speech
Evidence about perceptual characteristics of speech

Acoustic level

Represented by sequence of amplitudes of sound wave
Noise characteristics of room noise and mike noise
Limits of human speech signal

Figure A9.1. Levels, Their Representations and Sources of Knowledge



Initial — High probability of "Hello"
High probability of transition to Confused
Fair probability of transition to Request

New — "What can I ask?"
"Tell me something?"
"What do you know?"
Open up on synonyms
Expect many pauses

Request — Regular grammar
Three subgrammars
Request-System
Request-Job
Request-Resources

Answering — Possibly transitions to and from all states except Terminating.

Figure A9.2. Example of a Finite-State Diagram of State of Voice-CS User

	SENTENCE FORM				
English Request	COUNT	ATTRIBUTE	OBJECT	VALUE	Possible Response
1. How long has the system been running?	HOW-LONG		SYSTEM	RUNNING	The system has been running for three hours.
2. What DECtapes are on the system?	WHAT	DECTAPES	SYSTEM		DECtapes, DTA0, DTA1, DTA2 and DTA3 are on the system.
3. How many DECtapes are on the system?	HOW-MANY	DECTAPES	SYSTEM		There are four DECtapes on the system.
4. Are there two DECtapes available?		AVAILABLE	DECTAPES	TWO	Yes, there are two DECtapes available.
5. How many users are there?	HOW-MANY		USERS		There are currently seven users on the system.
6. Could you tell me what the null time of the system is?		NULLTIME	SYSTEM		We are currently running 37% null time.

Figure A9.3. Examples of Elementary Sentence Form: COUNT ATTRIBUTE(OBJECT)=VALUE

keyword analysis, a simplified technique of linguistic analysis in which only key words in the lexical stream are detected and much of the grammatical structure is ignored. This technique has been used in a number of applications (e.g., in the so-called ELIZA programs, Weizenbaum, 1966, 1969). It works quite well if the situation is sufficiently unsubtle. That this is the case can be seen in Figure A9.4, where an ELIZA-type program has been given a script for the computer task at hand. It provides a close correspondence to that in our protocols for the status task (Figure 5.1 or Appendix 6). In Voice-CS the keyword analysis technique is augmented by the contextual dependency analysis which is based on the constraints imposed by the keyword on the elementary sentence form. In absence of rigid syntactic structure, Voice-CS can still benefit from the order-free constraints imposed by the Voice-CS grammar after a keyword is recognized.

Lexical Level: The representations for the top two levels are the ones special to Voice-CS; the other levels have task independent representations. The lexical level is represented by a sequence of words. These words form a very limited dictionary (given in Appendix 6), and do not include all English words.

Phonemic Level: This level is represented basically by a sequence of phonemes.* The phonemic alphabet is one devised for Voice-CS, and is not identical with the standard alphabet of about 40 phonemes. An actual selection of this alphabet will not be made, but it will be an attempt to characterize the acoustic segments rather than a linguistic transcription. Actually, the phoneme representation consists of a sequence of lists of phonemes. The identification of a phoneme is sufficiently unreliable that, for each place (as indicated by a segmentation process), there needs to be a list of the likelihood of their being the actual phoneme.

As an alternative, there are many advantages to representing a phonemic segment by a simultaneously occurring set of attributes or features. Uncertainty in phonemic identity can be expressed by setting particular features to values intermediate between 0 (does not possess feature) and 1.0 (does possess feature). In the lexicon one can specify a range of values that each feature of each segment must fall within in order for an unknown sequence to match that lexical item. Thus, certain phonemes such as those falling in the stressed syllable can be emphasized by relaxing the range requirements in lexical representation of phonemes of unstressed syllables. A feature representation that is tied closely to articulatory variables such as the position of the tongue body, tongue tip, lips, etc., permits the decoding of coarticulatory phenomena that occur for example at word boundaries. Otherwise, the lexical entries will have to allow great latitude in the expected phonetic quality of word initial and word final phonemes due to coarticulation with adjacent words. The following discussion could be rephrased within a feature framework but for clarity we have chosen a single representation in terms of phoneme sets.

Parametric Level: The parametric representation consists of a set of measurements taken over time. Thus, if there are 8 measurements, then the parametric representation consists of 8 coordinated sequences of measurements. The actual set of parameters to be used will not be selected, but discussion of the possibilities and the consequences of various choices will occur later.

Acoustic Level: The final level consists of the representation of the speech signal as a function of time. Whether this exists as a digital sequence (after passing through an a/d converter) or exists only as an analog signal depends on the details of how the parametric values are derived (e.g., through a hardware preprocessor or through software). This may be an important decision in terms of the total processing capacity of the system, but is not relevant at this level of analysis.

We will illustrate the structure of the system by stepping through the system with an example from Voice-CS. Figure A9.5 shows the overall structure of the system for Voice-CS. Each of the levels again appears, with a brief reminder (in angle brackets, < >) of the representation. Beneath each level is a description of the mechanisms used at that level. This is not a regular flow diagram, since the actual sequence of control is not the most important aspect of Voice-CS, but rather what mechanisms use what knowledge to determine representations at what level.

Let us enumerate these mechanisms by stepping through the system with an input. Figure A9.6 shows the course of this processing, starting with

---

* There is, actually, no term available to describe the first division of the continuous speech into intervals identified with an element of a finite alphabet. Phoneme, phone, phonetic segment, hyper-phoneme, electroneme, etc., have all been used and each has specific associations. We will use phoneme without commitment to a particular phonemic theory or alphabet.

```
WOULD YOU LIKE INSTRUCTIONS FOR USING THIS PROGRAM?
*** Y
GOOD DAY.  I AM DR. OTTO MATIC.
SINCE WE CANNOT SEE EACH OTHER, THERE ARE A FEW CONVENTIONS I@D LIKE YO
U TO ABIDE BY IN ORDER FOR US TO COMMUNICATE AS WELL AS POSSIBLE.
     1) PLEASE WAIT FOR ME TO PRINT THREE ASTERISKS BEFORE YOU TYPE ANY
THING.
     2) TERMINATE EACH OF YOUR REMARKS TO ME WITH EITHER A QUESTION MAR
K OR A PERIOD.
     3) IF YOU MAKE A MISTAKE IN TYPING ENTER THE WORD
     4) WHEN YOU ARE FINISHED, A SIMPLE @GOODBYE.@ WILL TERMINATE THE I
NTERVIEW.
IF YOU WILL REMEMBER THESE RULES OUR SESSION SHOULD GO QUITE WELL.
NOW....
WHAT SHALL WE TALK ABOUT TODAY?

*** WHAT SYSTEM IS CURRENTLY OPERATING?
THE SYSTEM PRESENTLY RUNNING IS ADEPT.
*** IS ADEPT A TIME SHARING SYSTEM?
ADEPT IS PRESENTLY OPERATING.
*** WHEN DOES TIME SHARING OPERATE?
ADEPT IS PRESENTLY OPERATING.
*** WHEN DOES OS RUN DURING THE DAY?
ADEPT OPERATES DURING PRIME SHIFT AND OS IS ON THE AIR OTHER TIMES.
*** DOES THE SYSTEM SUPPORT BOTH DISC AND TAPE?
TWO TAPE DRIVES ARE AVAILABLE.
*** WHAT ABOUT DISC DRIVES, AND DO THE DRIVES USE PACKS OR ARE THEY
*** FIXED HEADS?
TWO DISC DRIVES ARE AVAILABLE
*** WHAT KIND OF DISC DRIVES ARE THEY?
I DO NOT UNDERSTAND THAT QUESTION.
*** WHAT MASS STORAGE DEVICES DO YOU SUPPORT?
THE SYSTEM SUPPORTS TAPES DISC DRUMS TELETYPES AND DISPLAYS.
*** WHAT IS THE SIZE OF THE MACHINE?
TRY AGAIN.
*** HOW MUCH CORE DO YOU HAVE?
THE COMPUTER HAS TWO HUNDRED THOUSAND BYTES OF MEMORY.
*** CAN I USE GRAPHICS TERMINALS ON THIS MACHINE?
THE SYSTEM SUPPORTS TELETYPES AND GRAPHICS TERMINALS.
*** HOW DO I USE THE DISPLAYS?
DISPLAYS APPEAR IDENTICAL TO TELETYPES IN THE SYSTEM.
*** IS THERE A FREE TERMINAL NOW?
I DO NOT UNDERSTAND THAT QUESTION.
*** CAN I LOG INTO THE SYSTEM NOW?
THE SYSTEM PRESENTLY RUNNING IS OS.
*** WHEN CAN I LOG INTO THE SYSTEM?
THE SYSTEM PRESENTLY RUNNING IS ADEPT.
*** CAN I LOG IN NOW?
PLEASE REPEAT THAT.
*** HOW MANY USERS ARE ON THE SYSTEM NOW?
SEVEN USERS ARE LOGGED IN.
*** IS THERE A LINE PRINTER?
THE LINE PRINTER IS NOT AVAILABLE.
*** WHEN WILL THE PRINTER BE AVAILABLE?
THE LINE PRINTER IS NOT AVAILABLE.
*** WILL THE PRINTER BE AVAILABLE TODAY?
THE LINE PRINTER IS NOT AVAILABLE.
*** FORGET THE LINE PRINTER?
I DO NOT KNOW WHAT YOU MEAN.
*** I SAID FORGET ALL ABOUT THE PRINTER.
THE LINE PRINTER IS AVAILABLE.
*** DID YOU SAY THE PRINTER WAS AVAILABLE?
THE LINE PRINTER IS NOT AVAILABLE.
*** GOODBYE.
MY PLEASURE.
```

Figure A9.4. Dialog from an ELIZA-like system simulating Voice-CS

Behavior level <sequence of responses, changes of user state>

Interpret elementary form
Make transition in user state system

Sentence level <sequence of elementary forms>

Sequence of key word parse programs
Sequence determined by user state

Lexical level <sequence of words from the dictionary>

Preprocessing of input sequence to improve order of alternatives
Using phonological rules
Using language statistics
Sequence of matches of phoneme strings
Search interval in input sequence determined by subgrammar and initial lexical pass
Subdictionary determined by subgrammar
Input phoneme string determined by pauses and reliable phonemes
Subset of subdictionary determined by reliable phonemes of input string
Initial pass to detect clear words, determines estimated user state
Reprocessing of close matches using additional parametric information

Phonemic level <sequence of phoneme-lists, where phoneme-list gives alternatives at point in order of likelihood>

Unitizing by continuity
Segmentation by discontinuity
Classification program

Parametric level <sequence of K parallel measurements>

Measurement programs for each parameter
Special measurements <especially for short intervals>
Under control of lexical level

Acoustic level <sequence of amplitudes>

Microphone pickup and digitizing

Figure A9.5. Mechanisms for Voice-CS Levels

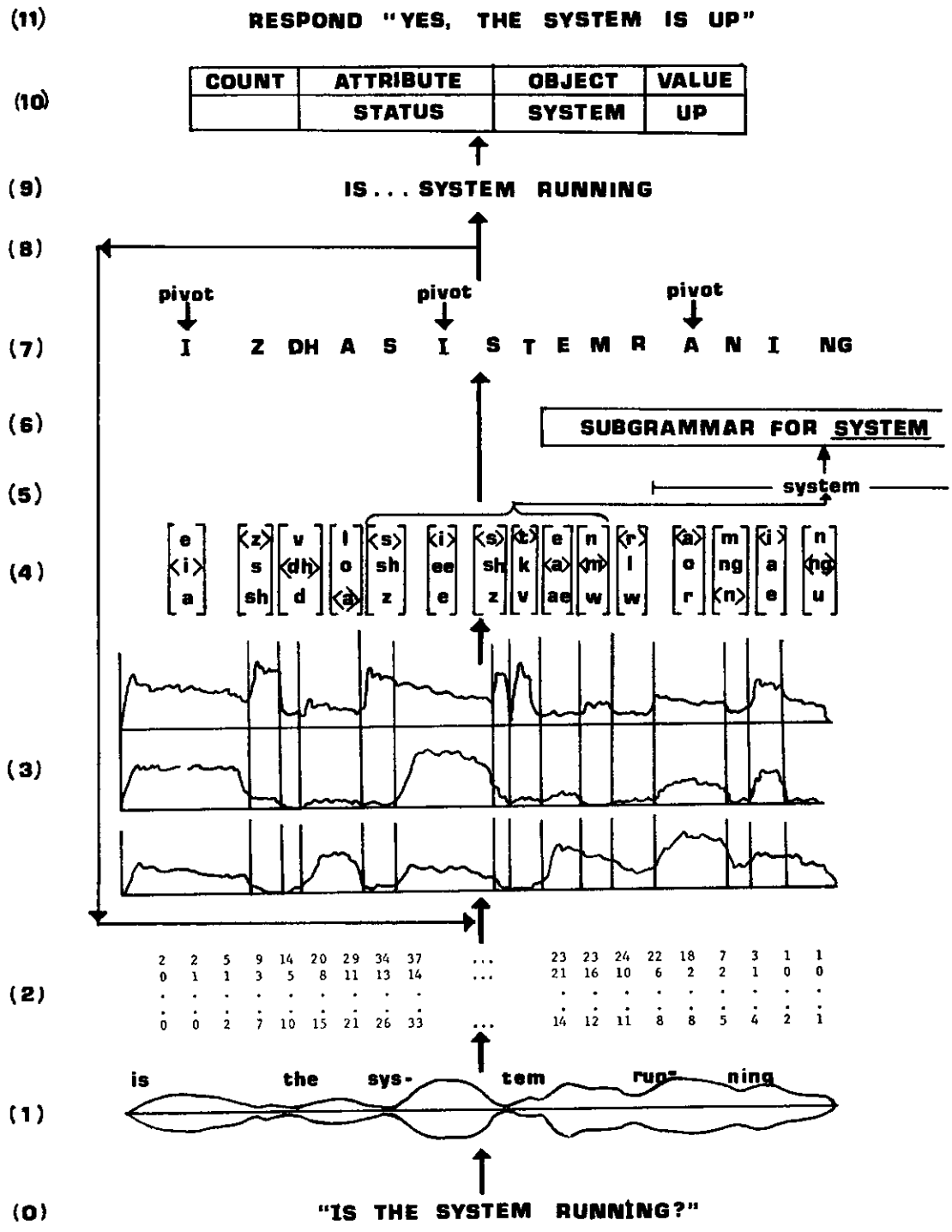


Figure A9.6: VOICE-CS PROCESSING OF AN UTTERANCE

the input utterance, written in English, at the bottom (0). This utterance arrives as an acoustic waveform (1) and is converted to a sequence of parameters (2). These are segmented by a process similar to that described for the Vicens-Reddy program, utilizing mostly information about continuity and variation of the parameters (3), after which a classification program operates to produce for each segment a list of phonemes (4). A pass is made through the sequence of phonemes against the whole dictionary. This is a stringent pass and only those parts of the input that are quite reliable are attended to. This yields (say) a single key word, "system" (5). This word is sufficient to indicate that the user is probably making a request about the system (as opposed to a particular job or a particular resource). Thus a new ordering is picked at the semantic level for what state the user is in. Each state has associated with it a particular limited key word grammar. In this case the grammar for the state request-system is selected (6). This subgrammar is given in Figure A9.7. Identification of other words (e.g., "hello") would have indicated a different user state (initial) and selected a different key word grammar. Thus, the state system shown in Figure A9.2, includes a separate state for requests associated with each object (system, resources, job, ...).

The key word subgrammar is now used to process all of the input stream in order to produce the lexical level. The central part of this process is a match between some interval of the input stream and some entry in the dictionary, i.e., the phonemic string corresponding to a word in the subgrammar (9). This match yields a quantitative estimate of how well the input segment matches the word, taking into account the entire list of alternative phonemes at each place as well as the possibilities that various phonemes were absorbed or totally obscured by their neighbors.

Conceptually, this match is to be iterated over the entire input utterance and over the entire dictionary in order to produce the representation of the lexical level, given in the figure as "is ... system running" (9). However, not all combinations are tried, nor are they tried in a fixed order. First, the relatively reliable features of the input stream are selected for accessing the dictionary. Thus, in the stream of 15 phoneme places that make up the utterance, only 3 access points occur (7). Somewhat broader limits are used than on the first pass, where only a single key word was needed to select the subgrammar. Each access point becomes a pivot around which a description is made that selects subpart of the subgrammar. That is, the subgrammar represents semantic selection and the subpart represent additional phonetic selection. Presumably, these are independent bases of selection. Iteration now occurs over all entries in this subpart.

A variant of the "empty world hypothesis" is now used: The speaker is probably talking only about one thing. If a few words are identified at a reasonable level of confidence which indicate a request that Voice-CS can understand, further processing is not worth much. Thus, the search starts with the high probability events and stops if it ever gets a complete message. Of course, it will sometimes make errors in so doing, which become one component of the total error rate of the system.

Since the system can stop looking when sufficiently good information has been accumulated, the order in which things are considered is important. Although no tight grammar exists, there are order effects among the key words. For example, if the key word is "how many," then almost for sure the other significant key words follow it in the utterance; if the key word is "available," mostly the other key words will precede it. This ordering mechanisms takes place early in the lexical level.

For each interval (more properly, each pivot point) there will be several candidates. Typically, one of two things happens. Possibly, one score is high enough, both absolutely and above its competitors, so that it can be taken as the word in the lexical string. Alternatively, none of the scores is high enough, compared to the statistical probability that an English word not in the subgrammar has occurred, so that the interval can be declared not to be a word in the subgrammar. The range in between these two is characterized by the existence of several leading candidates, whose scores are either too close or not high enough. Given the specific nature of the conflict (is it "magtape" or "DECtape"?), additional measurements are possible at the parametric level to make the decision. Thus, there is a feedback loop through the parametric, phonemic and lexical levels for such cases (8).

Given the lexical level, the sentence level is obtained by a table (the subgrammar of Figure A9.7), which assigns to each word in the vocabulary of the subgrammar the implications it carries for the elementary sentence form. The result for the example is shown at (10) in Figure A9.6. The various items of the elementary sentence form have been filled in. This is based on keyword analysis and contextual dependency analysis given the keyword. The semantic level is straightforward, which consists of interpreting this filled elementary sentence form to produce the response of Voice-CS to the input utterance (11).

In the example of Figure A9.6 the main path proved adequate, but of course this is not always so. The selection of a particular subgrammar was tentative, and if it does not work out then a different subgrammar must be tried. More information is available on which to select the next

COUNT	ATTRIBUTE	(OBJECT)	=	VALUE
	AVAILABLE			AVAILABLE
	BUSY			BUSY
		COMPUTER		
	CONTROL-C	CONTROL-C		CONTROL-C
	CORE	CORE		
		DATE		
	DOWN			DOWN
				EIGHT
				FIVE
				FOUR
	FREE			FREE
HOW-LONG				
HOW-MANY				
HOW-MUCH				
		JOBS		
				NINE
		NULLTIME		
				ON
				ONE
		PDP10		
		PEOPLE		
		RESOURCES		
				SEVEN
				SIX
	SIZE			
		SPACE		
	STATE	STATE		
	STATUS	STATUS		
	SWAPPING			
		SYSTEM		
				THREE
		TIME		
				TWO
	UP			UP
		UPTIME		
		USERS		
	VIRTUAL			
	WORKING			WORKING
				ZERO

Other words: ARE, HOURS, IS, K, MINUTES, NOT, SECONDS, WHAT, WHERE, WHO.

Figure A9.7. Request-System Subgrammar with Implications for Elementary Sentence Form

subgrammar, since some words will have high enough scores to stand, even though other candidates come on the scene. Thus, a new selection of a subgrammer can be made and the process starting from the lexical level can be iterated. This takes place against the background of the prior analysis, so that not all processing needs to be repeated and lower bounds exist for many phoneme intervals about how good the match must be to be taken as the recognized item.

Unsolved Problems of System Organization: Voice-CS system raises several interesting problems in the organization of the large computer programs that would be almost impossible to attempt on most present day computers. Any speech understanding system which attempts to include all the sources of knowledge we have outlined in the preceding pages will certainly be a large program with severe real-time requirements. Most presently available operating systems do not provide the necessary facilities for the development of such a system.

Voice-CS attempts to answer questions about computer status over a telephone. If it takes more than a few seconds to respond to a trivial question then the user would soon become disinterested in the system. It follows that to be acceptable Voice-CS must respond to trivial questions as soon as the question is completed. Indeed, to equal human performance, it must sometimes be able to answer questions even before they are completed. This means every mechanism must do its part as soon as it is able to.

As soon as the signal parameters are extracted the segmentation program must begin determination of acceptable segment boundaries. As segments become available, the phoneme recognition program must assign labels with associated probabilities. As a syllabic nucleus is formed with an apparent stressed vowel the keyword linguitic program must be activated. As possible candidates are formed the lexical recognition program must determine the most likely word just uttered. The availability of the keyword should then generate possible hypotheses for other words that might occur in this utterance, their possible location, and their relationship to the keyword already recognized. As more words are recognized the system should be able to decide when it has information to begin answering the question without waiting for the utterance to be completed.

At first look the subroutine mechanism commonly available in most programming languages might appear to be sufficient. However notice that the program is going from routine to routine in seemingly random order which is purely dependent on the data that is arriving from the microphone. This can only be done if each subroutine has facilities to interrupt its processing in midstream to preserve its state, and transfer control to another routine. When the interruption occurs at predictable points the co-routine concept commonly used in compiler writing would be sufficient for our requirement, i.e., preserve its state and continue upon return without any fixed entry-point initialization resetting the state.

In the case of Voice-CS, the co-routine structure proves to be inadequate, because the active routine can only relinquish control at preprogrammed points. If in the meantime the speaker has uttered several words that require immediate processing because of overflowing buffers it would mean irrevocable loss of data.

A parallel program organization is indicated. A time sharing does essentially what is needed except most present systems do not provide facilities for interprogram interaction. Most time sharing systems are built with the assumption that the programs under its control are performing independent tasks and do not need to interact with each other. In addition the system must guarantee service to some programs every 100 ms or so (the real-time problem). Very few present systems provide these facilities and this is likely to be a major obstacle in the immediate realization of demonstrable speech-understanding systems.

A9.2 Semantic Level

In this section we will discuss the problems and the sources of knowledge at the semantic and post-semantic levels. Many of the issues raised at this level deal with man-machine interaction: the new word problem, the synonym problem, the verification problem, the real time problem and the user model problem. To study some of these issues, a simulation of Voice-KP was performed by Dr. Robert Anderson of the RAND Corporation at the request of the committee. We will summarize the results of the simulation here.*

A9.2.1 Simulation of Voice-KP. The simulation of Voice-KP was conducted to answer two main questions: how fast can a person read-in routine specially formatted data and what is the effect of errors on the data rate? The first question is relevant because in normal conversation the semantic context of the conversation permits a great deal of sloppiness in both the speaker and the listener without any loss of the essential message. Further in normal conversation the mind can and does formulate utterances long before they are to be uttered. The question then is what happens to the data rate of the speaker when he is deprived of his language and environmental context? This is important for Voice-KP because much of the data that is keypunched on cards is formatted and

* We wish to express our deep appreciation to R. H. Anderson for obtaining the results presented here.

usually communicated among humans in written form and not by voice. Does this affect a person's ability to communicate it by voice at a high data rate?

The second question on the effect of errors on human performance is also important because almost any system that can be conceivably built in the near future is likely to be errorful. If the speaker has to frequently correct the data there will be a threshold of patience at which he will prefer to use some other medium. The questions to be answered are what error rate will he tolerate and can it be obtained within the present state of the art?

The System: The performance of a Voice-KP system was simulated by a man-machine system. The man simulated the recognition and interpretation part of the Voice-KP listener. The machine simulated the errorful behavior and the visual feedback parts of the Voice-KP system. After recognition and interpretation the listener of the Voice-KP does not have enough time to type-in what the speaker said (this would artificially slow down the data rate of the speaker making the results meaningless). So it was necessary to preprogram in the data to be read by the speaker ahead of time and all that the listener had to do was hit a key soon after the word was uttered. Since the anticipatory reaction time of a human being is in the range of 200 to 400 ms the delay was not considered unreasonable. Detailed description of the system setup is given in Appendix 5.

The Experiment: Several lines of data of the following format were read-in by the speaker:

Emp. No.	Surname	Initial	Sex	Age	Marstat
00365	Charleston	G.	M	37	M

Dep	Draft	Degree	Major
0	3A	MA	Math

Complete listing of the data used is given in Appendix 5.

The subject's speech was restricted to the following format: he can mention data values in order from left to right, or he can reposition himself at any column by mentioning the column heading. In addition, he had two control commands RESET to start him over again and NEXT to proceed to the next line of data.

Results: Approximate average times for data input was 4 minutes at 6% error rate, 5 minutes at 12% error rate and 6 minutes at 24% error. It took approximately 2 minutes for 0% error rate without waiting for visual feedback (this represents the situation when the user knows he is dealing with a perfect recognizer) and 3 minutes for 0% error rate with visual feedback. The time for typing and proof reading the same data was 4.5 minutes and the time for keypunching and verification on cards was 5.5 minutes.

Conclusions: The results presented above are based on very limited experimentation and thus these conclusions must be regarded as tentative. The answers to other questions we raised seem to be 1) on an error free system the spoken data rate of non-speech like material without verification is about 3 times as fast as keypunching and 2 times as fast with verification, 2) on an errorful system the performance of Voice-KP over keypunching is slightly better with less than 10% error rates and slightly worse when the error rate is greater than 20%. These results show that Voice-KP is an attractive alternative to keypunching only when the data is self generated. A Voice-KP system might be useful at the data source where some one must transcribe his data onto a paper which can then be keypunched, e.g., reading of gas and electric meters by the utility companies or inventory taking at a warehouse.

A9.2.2 The Real-time Problem. A crucial problem for Voice-KP is that the system must keep up with the data rate of the speaker. Unlike the other three tasks where a pause of a second or two after the question might go unnoticed, the Voice-KP system would lose data if it cannot keep up with the speaker. We have discussed this problem in part in A9.1 under the system organization problem. The requirement here is more severe. It is not sufficient to have a program which will use disk-spooling to keep the data from getting lost, but rather it becomes necessary that every mechanism perform its function so that the sum total of the time is less than the time taken to utter the statement.

This requirement has grave implications. It is often the case that factors of 3 to 10 improvement in performance of a system are only achieved by new breakthroughs. Breakthroughs often depend on (or require) radical departures from the known existing solutions. This means after building a system that works satisfactorily for Voice-CS or Voice-DM, we may find that this system cannot satisfy the real-time requirement of Voice-KP and force the development of a completely new system. Clearly some of the basic mechanisms will still be useful but the total system would have to be programmed anew.

A9.2.3 The New Word Problem. How does one add a new lexical item to the data base or specify that a new format of input is to be used by means of voice? At one extreme one could spell the word using international spelling alphabet of "alpha, bravo, charlie, ..." each time a difficulty is encountered. Or the system could have a model which guesses the word spelling from the rules of orthography of the English language given an

approximate phonemic transcription of the word. The latter technique with facilities for prompting and error correction would probably be desirable. Whether such a system can be built to operate satisfactorily can only be answered by further research. It is known however that the international spelling alphabet can be recognized highly accurately for single speaker systems (Bobrow and Klatt, 1968). Whether a system capable of recognition of this alphabet for most native speakers can be built is also a question to be answered by future research.

Specification of a new format for input raises other problems. A new format of the type used in Voice-KP simulation merely reduces to the preceding problem of learning new words and their spelling. However, if the format is specified by means of, say, a phrase structure grammar of the input language, then we face all the problems of extendable languages that are being faced by the researchers in the programming language field (Galler and Perlis, 1966; Cheatham, 1966; Standish, 1969). Whether it is feasible or desirable to effect such modifications by voice is questionable at this point in time. In the case of Voice-DM the solution to the new problem can be substantially simplified. Instead of voice input a keyboard can be used by an expert for the addition, deletion and modification to the data base. He would not only type in the phoneme transcription of each new word entered, but would also modify the data representation so that a new word would be entered under the appropriate category within the data structure of the phonemic lexicon. The problems associated with the representation of the large lexicon will be discussed in Section A9.4 under the lexical level.

A9.2.4 The Synonym Problem. When faced with the choice of several possible synonyms, the actual one chosen by the speaker appears to be random. A close study of the protocols taken in conjunction with the simulation of Voice-KP shows that the same speaker within the same list has used "M", "Male", and "Sex Male" in similar situations. Other examples include "Math" or "Mathematics" and "Age thirty-one" or "Thirty-one" or "Three one". The stranger the data the greater seems the synonym variability.

The implications of the synonym problem for Voice-KP systems are clear. All the synonyms must be anticipated and programmed. It would be desirable if a program could be built for the automatic generation of commonly occurring synonyms from a study of human synonym usage.

A9.2.5 The Verification Problem. Using the same sensors to perform two competing tasks leads to several unanticipated problems. In the case of the simulation of Voice-KP, the speakers eyes were moving continuously from data to the display and back. Not only was this strenuous, but proved to be time consuming. It was estimated that it took anywhere from 5 to 10 eye movements for the location and registration of a data element in the midst of utterance.

If eyes are to be used for the data generation, then it seems necessary to use some other sensor for verification. That we need verification for Voice-KP is clear from the task. Verification by means of voice response from the system seems to be the main alternative. Other sensory mechanisms are not capable of the verification that is needed in this case. The role of the eyes and ears could be switched if needed, e.g., if the speaker is transforming voice data into a form suitable for Voice-KP, then eyes could be used for verification. The verification problem thus introduces previously unanticipated design constraints on Voice-KP.

A9.2.6 The User Modeling Problem. The total semantic situation is constituted of a number of distinct components, each of which can make a contribution to narrowing the possibilities for what is said on a given occasion. The main one (sometimes even identified as the semantics) is the structure of the task environment. It forms a sequence of increasingly particular contexts. For example, using Voice-CC as an instance, we might have:

(1) Becoming familiar with the new Tenex monitor (the original Voice-CC task)

(2) Understanding how to use the file system in Tenex (a particular subcontext)

(3) Reacting to the PDP10 response of "?" to an attempt to give a file command (an immediate context).

Each of these contexts increasingly restricts the plausible responses. Words, such as "file" and "can't" become more probable in context (3) than in the task as a whole. And these restrictions can be derived in part from an objective analysis of the task to be done.

The restrictions indicated above can be derived largely from an objective analysis of the task to be done. But there remains appreciable freedom for human action and, in turn, appreciable contribution from the psychology of the user. For instance, still considering the above situation, the following strategies are all plausible:

(1) Distrust prior action, hence repeat same action.

(2) Guess at command, attempt variation.

(3) Look in manual.

(4) Ask specific question to obtain correct command.

(5) Make general appeal for help.

(6) Go on to another task and come back to this one later.

The strategies are characteristic of humans of a particular general character (namely, educated adults at home in technical society). Given the above options it is possible to be much more specific about what might be said. Even more specificity can be obtained if we know of a particular user that he almost always follows the strategy:

(7) Do (2) for several trials and then do (5).

Even though we do not have a specific figure for the number of trials, we can give the vocabulary for strategy (2) first priority for (say) two trials and then switch to the vocabulary for strategy (5) from there on. In all events we can put much lower priority on the vocabularies of the other strategies.

No studies are available on the Voice-CC task to show whether humans of a given general type limit themselves to a finite set of strategies (e.g., (1) - (6) above), or whether individuals are consistent in their selection of a particular strategy from this set (e.g., (7) above). However, in some analagous tasks (puzzle solving and game playing) such consistency does occur (Newell and Simon, 1971) and we would expect it in the Voice-CC task. Immensely detailed study would be required, but such models of psychology of the individual user seem possible. (They are almost certainly used by one individual in understanding another, since the hearer can essentially put himself in the position of the speaker and ask what he would do in the situation, all without excessive cognitive strain).

The types of information given above are still essentially semantic, i.e., they pertain to the meaning of the sentences to be uttered, not to the actual choice of words and phraseology. Given that a user decides on strategy (3), he can still say:

(1) (nothing)

(2) "Well, let's look at the manual see what's up."

(3) "O.k., look it up."

(4) "All right, to the manual."

(5) "Wonder what page that would be on."

...

Although the variability is still high, so that no single word (not even "manual") is common to all sentences, substantial constraint has been introduced.

Almost no information appears to be available on how variable such task-oriented, but freely-emitted, speech is. There are undoubtedly conventions of discourse that are generally observed by all people (of a given general type) and more specific stylistic consistencies within an individual.

An additional source of knowledge comes from the known basic characteristics of the human as an information processor. A human can remember only a few things at once, without adopting a deliberate strategy of rehearsal; similarly he can acquire new information only at a relative slow pace (a few items per minute). These provide a basis both for predicting what the subject will not say, and when he will ask for a repetition of information (if too much intervened). Another regularity of great importance is the fact that humans operate with a goal stack. When interrupted, they do not abandon what they were doing, but save it while they take care of the new problem, then return to it and attempt to pick up at the point of interruption. In any few minutes of task-oriented behavior (such as occurs in Voice-CC) several minor interruptions will occur (e.g., finding a little used key on the teletype, picking up a dropped pencil, recalling the meaning of a computer-sent message, etc.). By keeping track of the tasks the human is working on, i.e., simulating his stack of goals, a substantial increase in predictability is possible.

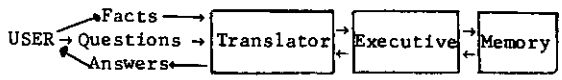
The purpose of the above discussion is to emphasize that a part of the knowledge available in understanding a speech utterance is in a psychological model of the speaker, rather than in the structure of the task (or in the various levels of linguistic structure). Little has been done to develop models of the speaker that would be genuinely useful to a speech-understanding system, but a few of the possibilities are indicated above.

A9.2.7 General Semantics. Unlike the other three tasks considered, Voice-CC requires the use of powerful semantics for its success. In the long run, this task and its model of general semantics offers the main hope of a general approach to semantic information processing and question-answering. All of the prior three examples used little or no semantics and what little there was could be handled by ad hoc methods for semantic representation.

State-of-the-Art in Semantics and Question Answering. Many of the important papers relevant to our work have been collected in Semantic Information Processing (Minsky, 1969). Two surveys of the state of the art are given

by Simmons (1965, 1970). A detailed characterization of the structure of question-answering systems can be found in Green (1969). We present here a brief description of the characteristics of question-answering systems. A question-answering system may be broadly defined as a system that accepts information and uses this information to answer questions. Often the information, questions, and answers are presented in a form that is relatively easy for people to learn, such as some restricted class of type-written English sentences. If the question-answering system, a computer program, produces reasonable responses, it may be attributed to the human characteristic, "understanding."

The following diagram shows the essential components of a question-answering system.



Its operation is as follows: The user presents statements (facts and questions). A translator converts them into an internal form. Facts are stored in memory. (The store of facts is referred to as the data base.) Answers to questions are formed in two ways: (1) the explicit answer is found in memory, or (2) the answer is computed from the information stored in memory. The executive program controls the process of storing information, finding information, and computing answers.

A question-answering system does not explicitly store all information that is available to the user. Instead, a smaller data base of compactly coded facts is used. New information, not explicitly stored in the data base, but implied by the stored facts, is computed or deduced from this data base by an answer-computation mechanism.

Semantic Constraints on Phonetic Analysis. The semantic support which one can expect from a semantic data base falls into two classes--general semantic information such as semantic selectional restrictions between verbs and their objects, subjects, and modifiers, and specific factual information about the world or about a data base. An example of the former might be the restriction of subjects of the verb "fly" to birds, insects and airplanes (in the literal sense of "fly") and to people (in the sense of being carried by a plane). An example of the latter would be the specific fact that the DODO bird can't fly, or that a particular airplane flight has been cancelled on a particular day. In the first case, semantic support for the speech recognition task is similar to the syntactic support discussed earlier. In sufficiently limited artificial languages this type of constraint can be included in the syntactic categories of the grammar--making it essentially a syntactic constraint. In an application where the semantics are less limited, general semantic information can still be used to predict words which might occur in some environment. A semantic association network such as Quillian's TLC program (Quillian, 1969) would be a good candidate mechanism for implementing this type of semantic support. Moreover, this type of semantic word association in predicitive mode is likely to be more useful in general English than the corresponding syntactic support since it imposes considerably more restriction on the predictions than does the prediction of a syntactic category.

In addition to the proposal of specific words, general semantic information can also be used to screen the tentative words proposed by the word recognizer in the same way that a grammar can be used to screen tentative words. The use of specific factual information is somewhat more difficult. It is not likely that the use of specific factual information to propose words will be of much help to the word recognizer, since the number of specific entities in a data base that can be referred to in a given environment may be quite large. Thus, for example, in a context where one is expecting a proper noun, an enumeration of all of the names in a personnel file would be of little help in proposing candidate words for the word recognizer. On the other hand, specific factual information could be very useful in resolving residual ambiguities in word recognition. Consider, for example, a case in which the word recognizer finds two possible flights is already on the ground. This type of semantic support, however, requires the use of some fairly powerful inferences in general, and the state-of-the-art of mechanical inference is still quite limited.

Sources of Error and Knowledge.* The range of questions illustrated by the sample protocol in Appendix 7 steps somewhat beyond the bounds of existing Q. A. systems by asking "How do I . . ." and "What happens if . . ." questions. In addition there are a number of questions (such as #11 and #27) which either are too much to expect of a Q.A. system (with today's state-of-the-art) or are not really necessary. Thus, in order to build a system with current (or almost current) state-of-the-art question answering it will be necessary to restrict the semantic range of questions somewhat from that represented in the sample protocol.

These restrictions can either be presented to the speaker in advance; or he can receive feedback of the type "I don't understand your usage of XXXXX, please rephrase;" or a combination of both. The restrictions would exclude utilization

* The restrictions and the protocol in Figure A9.8 were proposed by J. Corbonell of BBN.

of syntactic and semantic techniques not yet developed and tested or not derivable as natural simple extensions of current work.

One limitation should refer to anaphoric references, and especially ellipsis which should be avoided. Uses of "that," "this," "it," etc. should also be restricted. For example, "it" in statement 23 of the protocol in Appendix 7 refers to "command," and not to "space" or "file," creating an anaphoric reference difficult to solve. Of course, we could give the computer the capability of asking the user something like: "What does your word 'it' stand for?"

Another possible limitation relates to poor syntax. Incomplete and other syntactically incorrect sentences are common in verbal discourse and can be observed throughout our protocol. One would like to use semantic information and phonology (e.g., intonation as a question for an otherwise affirmative sentence) to force interpretations of ungrammatical sentences so as not to force the user to express himself in prefect English. However, except for the ELIZA-type approach, this task is still frontier state-of-the-art.

Most types of "if" clauses and other conditionals should be avoided since they create complex semantic problems. From a syntactic point of view, involved conjunction or disjunctive sentences can easily generate complex ambiguities, the solution of which is at best a time-consuming process. Limitation of conjunctions to the simplest cases seems advisable. In general, one should instruct the man to employ, as much as possible, short, simple, and direct sentences avoiding complex and involved statements.

One aspect of semantics that will require further elaboration beyond current state-of-the-art is that of processing actions and commands. Most current Q. A. systems are not mainly concerned with questions of "doing." They handle well concepts and facts, with less emphasis on procedures. Questions like "How do I copy an existing file from disk to tape?" refer to actions. This extension hopefully will be relatively straightforward, and work is already in progress towards it.

Figure A9.8 presents a brief simulated protocol of a possible man-computer verbal exchange taking into account the limitations expressed above.

* * * * *

User: Can I copy files?

Computer: Yes, you can copy files.

User: What's the procedure I should use if I wished to copy an old file?

Computer: Please rephrase your question. I don't understand it.

User: How do I copy an old file onto a new file?

Computer: You can use the copy command which is of the form copy file to new file. Why don't you try it?

User: (types) What does the period mean?

Computer: Do you mean the period I added to your input?

User: Yes.

Computer: A period normally separates the file name from the file extension. In this case your file has a null extension.

User: Is the version number of a file the same as the extension?

Computer: No, they are different things. The format is: file name, period, file extension, semicolon version number.

User: I didn't specify the version number above and the system didn't complain.

Computer: What is your question please?

User: If I don't specify the version number of a file, what will the system do?

Computer: Nothing.

User: Explain, please.

Computer: TENEX will automatically insert the proper version number for you. If you are writing a new file, TENEX will consider it as version number one.

User: What if it's an old file?

Computer: Your articulation is unclear to me. Do you mean to ask what an old file is? Repeat please.

User: What happens when the written file is old?

Computer: When the written file is old, if the file version is unspecified, TENEX adds one to the last version number.

Figure A9.8. A simulated protocol illustrating the restrictions for Voice-CC

A9.3 Sentence Level

In this section we will review the art of syntax analysis to show that the techniques available are adequate to the task provided the string is error free, discuss new problems that arise in parsing speech instead of written text, discuss the nature of constraints provided by the sentence level, and present some possible mechanisms for interfacing syntax to the lower levels.

A9.3.1 Review of the State of the Art. Stockwell, Schachter, and Partee (1968) have been gathering the currently available information on transformational grammars for English, and the bibliography of that paper contains an extensive coverage of what is currently known about the English language in terms of transformational theory. In addition, there exists considerable knowledge about parsing algorithms, and formal grammars for natural language. Kuno (1967) gives a summary account of the use of computers for natural language analysis.

Available parsing procedures for natural language include the original Harvard Predictive Analyzer of Kuno and Oettinger (1962), MITRE English preprocessor (1964) and Petrick's (1965) analysis procedure for transformational grammar based on Chomsky's theory of transformational grammar, and more recently, various transition network grammars (Thorne, 1968, et al., Bobrow and Fraser (1969), and Woods, 1969), Winograd's (1971) procedural grammar, Martin Kay's (1964) "powerful parser," versions of which have been used by Kay, Simmons, and others for the parsing of natural language. Of these, the Predictive Analyzer applies only to context free grammars and is inadequate for characterizing the subtleties of natural language, although it is still the largest machine grammar for natural language that has yet been implemented. The Petrick analysis procedure for transformational grammars is too exhorbitantly slow for any practical applications involving the parsing of a large number of sentences. Potentially useful parsers are the "powerful parser" of Kay, and the augmented transition network model of Woods and Winograd's procedural grammar.

Of these, the augmented network model and the procedural grammar model are the most recent and appear to have many characteristics which make them especially well suited for the speech analysis application--notably, the presence of real syntactic hypotheses about the sentence at the time when the next word is being isolated from the input string.

Parsability of Verbal Statements. In order to assess the difficulty of parsing the rather unrestricted questions of the sample protocol, we attempted to parse about a dozen of them with an existing grammar for a fairly large subset of English. This grammar is an augmented transition network grammar for the NET2 experimental parsing system at BBN (Woods, 1969). The only initial changes made consisted of setting up dictionary entries for the "words":

ah

er

um

. . . (pause)

indicating that they should be skipped. The new vocabulary words were added as needed. The experiment brought to light a number of arcs that were missing from the grammar used, but no problems of any appreciable difficulty arose. Typically, when a sentence failed to parse, it was merely necessary to add one new arc for a constituent that was not handled in the original grammar (e.g., an initial adverb beginning a sentence) or to adjust a condition on an arc. It does not appear to be appreciably difficult (at least with a transition network grammar) to make a grammar for the range of utterances illustrated in this protocol (except for 1 or 2 totally ungrammatical ones). (One advantage of the transition network grammar is the ease of backing up and ignoring false starts that occur somewhere in the middle of a sentence.)

The sentences uttered by the user in the sample protocol are almost all quite simple and fairly short. This seems to be somewhat a characteristic of the task, although the sample is based on only one user and there is probably considerable variability among users in this respect. Nevertheless, the syntactic parsing of at least a habitable subset of "speech-type" utterances (i.e. utterances including false starts, "ah's," "er's," pauses, etc.) is clearly feasible.

Figure A9.9 gives some examples of syntax analysis of sentences from the Voice-CC grammar using the augmented transition network model of Woods (1970). Figure A9.10 gives an overall flowchart of the syntax analysis program of Woods to illustrate the structure of syntax analysis program.

In addition to these models of natural language, it is possible for many limited applications to define an artificial subset of natural language which admits a purely context-free grammar. In this case, there are many context-free parsing algorithms to choose from. One of the most efficient is the context-free recognition algorithm of Jay Early (1970) which operates within the best known theoretical time bounds for every subclass of context-free grammar for which time bounds are known. Moreover, this algorithm can be applied to unaugmented transition network grammars with an additional improvement over ordinary context-free grammars.

```
**(HOW DO I LOG ONTO TENEX?)
PARSINGS:
S Q
  NP PRO I
  AUX TNS PRESENT
  VP V LOG
     PP PREP ONTO
        NP NPR TENEX
     ADV HOW
**(NOW WHAT DO I DO?)
PARSINGS:
S Q
  NP PRO I
  AUX TNS PRESENT
  VP V DO
     MP DET WHQ
        N THING
        NU SG/PL
     ADV NOW

**(OKAY)
PARSINGS:
S EXPL
  OKAY

**(OKAY, AH, ... DO I HAVE TO , AH, PUT A PERIOD OR ANYTHING?)
PARSINGS:
S Q
 NP PRO I
  AUX TNS PRESENT
  VP V HAVE
     NP NOM
        S FOR-TO
          NP PRO I
          AUX TNS PRESENT
          VP V PUT
             NP OR
                NP DET A
                   N PERIOD
                   NU SG
                NP DET ANY
                   N THING
                   NU SG

**(OKAY WHAT'S THE PASSWORD)
PARSINGS:
S Q
  NP DET THE
     N PASSWORD
     NU SG
  AUX TNS PRES
  VP V BE
     MP DET WHQ
        N THING
        NU SP/PL
```

Figure A9.9. Examples of Augmented Transition Network Syntax Analysis

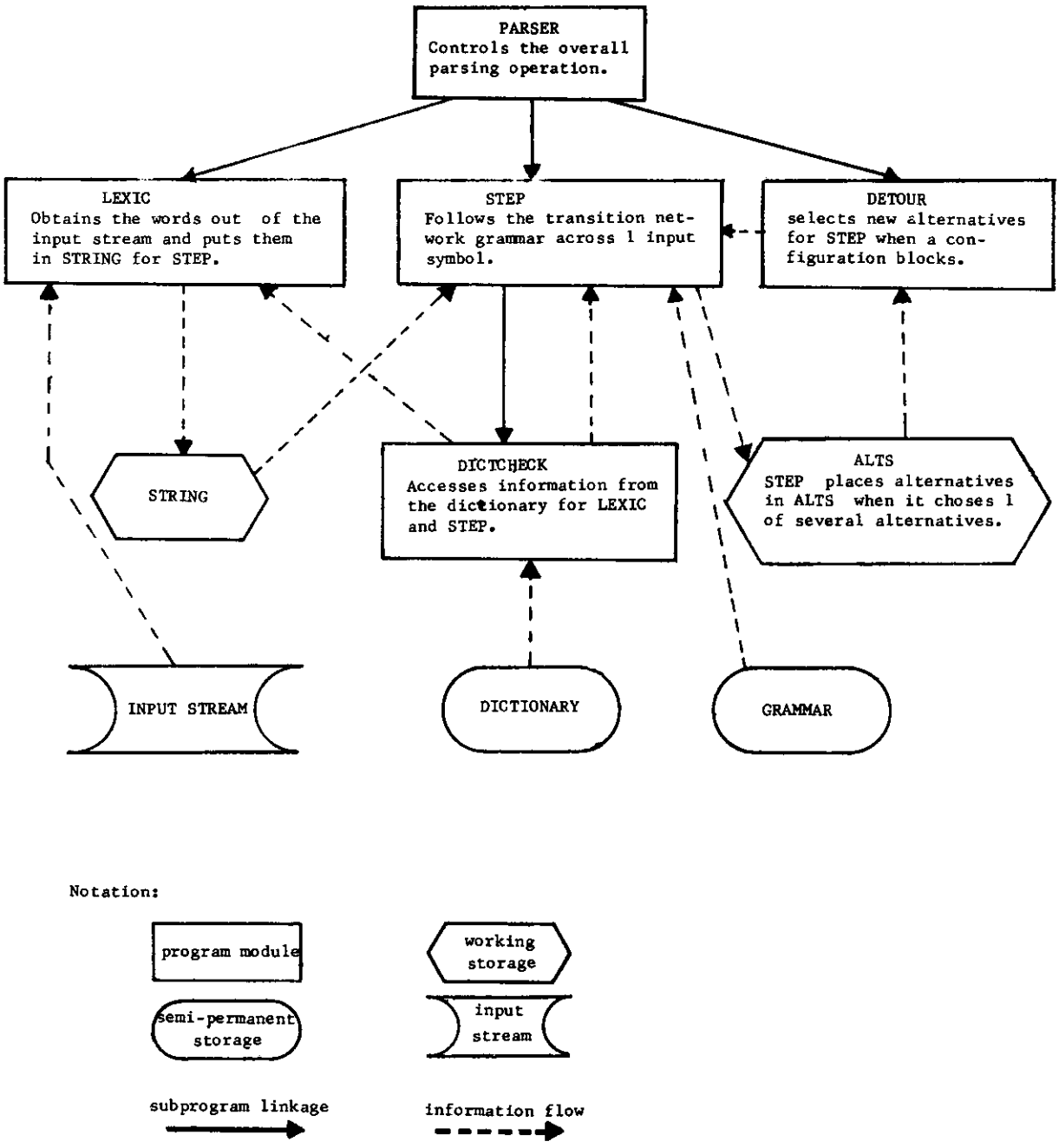


Figure A9.10 Flowchart of Augmented Network Syntax Analysis Program of Woods.

A9.3.2 Unsolved Problems at the Sentence Level. The non-grammaticality and non-wellformedness of speech utterances present several problems that are peculiar to speech. In addition, there are no word boundary indicators in speech analagous to "space" in the written text.

Parsing in the Presence of Errors. Acoustic characteristics of function words (usually unstressed) within a speech utterance tend to exhibit wide variability. This makes it difficult to use the conventional parsing techniques in the analysis of speech utterances. What seems to be needed are parsers which can parse both backwards and forwards, starting from anchor points in the middle of the utterance (content words which are usually stressed). Very little work has been done in this direction. At present it is not clear whether there are some simple ways of adapting the existing parsers so that they can handle this situation adequately or whether new techniques have to be developed.

Parsing in the Presence of Noise. We have already discussed some simple ways of ignoring noise words, such as "ah," "er," "um," within the framework of presently available parsing systems. However, there still is the problem of being able to predict that some part of the acoustic signal, in fact, represents a noise word. In addition to the noise generated by the speaker, one has to consider the problems associated with parsing of sentences in the presence of external noise, such as background clicks, laughter, or other speech.

Interpretation of Partial Parses. Often people have a tendency to abbreviate sentences leaving out the whole subject or predicate. For example: the partial parses that would result from the protocol in Appendix 6 -- "(Laughter) (b) tough...ummm...ok. That is all I want to know...Period. And handing up will cut off this conversation." It appears necessary that systems should be able to recognize that parsing in its present state should be suspended and a new parse attempted of the remaining utterance and then at some point re-evaluate all the partial parses to see if they should be ignored or combined to obtain appropriate interpretations.

A9.3.3 Constraints Provided by the Sentence Level.

Constraints for Voice-DM. A main source of knowledge for Voice-DM is its highly restricted syntax. The evidence from Vicens-Reddy system shows that if you restrict the syntax sufficiently then it eliminates the need for all other sources of knowledge (although one may end up with a trivial language). The expected reduction in search and disambiguation obtainable from the syntactic constraints of Voice-DM have not been estimated. Preliminary calculations show that the reduction will only be of the order of two or three and not of the order of ten or more.

Only the words known to the system will be used for querying the Voice-DM system. This absence of lexical noise makes it easier for the system to choose a word with a low similarity score provided it is greater than all other scores. In Voice-CS, a word with low similarity score may not be accepted if it is below a threshold since it may be one of the words that is not part of the acceptable vocabulary. The absence of external noise resulting from the use of a high quality microphone in a quiet room permits Voice-DM to make finer disambiguation than is possible with Voice-CS.

Constraints for Voice-KP. The restricted format of the input data provides significant constraints on the vocabulary. For example, if the next item to be spoken is sex, then the only meaningful words are "M," "F," "Male," or "Female." If none of these receive a high enough similarity score to the incoming utterance, then one would consider the possibility of one of the format words like "Sex," "Age" or the control words "Reset" or "Next." In many cases this could mean more than a 90% reduction in the search of the lexicon.

Highly constrained format like the one used in the simulation has another beneficial effect. Most speakers seem to pause involuntarily at various category boundaries. This results in separated speech which significantly reduces the lexical segmentation ambiguity and search. High quality microphone input and the ability to correct errors immediately permits much larger vocabularies to be handled by Voice-KP than is possible for Voice-CS.

Constraints for Voice-CS. In Voice-CS the user is not required to follow any rigid syntax. However, the fact that he will ask questions in English and the fact that he is dealing with a very limited task domain imposes several contextual constraints once a keyword is recognized within the utterance. This source of knowledge can be used either to restrict the lexicon match to a few words or to order the possible candidates from the lexicon.

For example, given that the keyword "job" is recognized as part of the utterance and assuming that the user is knowledgeable and cooperative (he deserves what he gets if he is not), it has to be a question about the status of the job, devices and resources being used by the job, the name of the user running the job, and so on. This implies the most likely keywords to be found in the utterance are "status," "teletype," "user," "program," "size," "routine," "resources," "running," "I-O-Wait," and so on. That is, only a small subset of the total vocabulary is induced. Conversely, if one of these other words has just been recognized as a possible keyword then it is clear that "job" would be one of the more likely keywords associated with it.

An accurate determination of the reduction in effort obtainable by the use of this contextual dependency has not been made for the Voice-CS vocabulary. It appears to a function of the semantic power of the word just recognized. Some preliminary calculations indicate that the reduction in effort may range anywhere from 25% to 75%. As more keywords are recognized within the utterance contextual dependency provides greater subselection among the possible words to look for next.

Constraints for Voice-CC. An attempt was made to estimate the degree of syntactic constraint imposed on the words of a sentence by the existing transition network grammar. It is a difficult thing to measure, especially for this grammar, because the arcs that are permitted at a given choice point are generally a function of previous history and features of the current word (e.g. for person/number agreement between subject and verb). Moreover, the various alternatives at a given point in the string are not all tried at once, but they are tried in an order which usually tries the most likely choice first. Thus when the parser is trying to determine the word, it can be doing so on the basis of a much more restrictive hypothesis than just the totality of words that can occur at that point. For example, a given state may accept either a relative pronoun, a preposition, or a noun in that order of likelihood. The parser would first ask if the current word is a relative pronoun (very restrictive hypothesis) and if successful would suspend the other alternatives (to be tried later if the current choice doesn't work out). If the word was not a relative pronoun, then the parser would ask if it could be a preposition (again a fairly restrictive hypothesis) and so on.

Bearing the preceding qualifications in mind, the following things can be said about the degree of syntactic restriction imposed by a transition network grammar. Out of 41 states in the network, 16 of them take only unconditional actions (pushes or pops) which neither depend on the input word nor move the pointer in the input string. These states, therefore, do not effect the syntactic constraints on the phonetic analysis nor do similar arcs leaving other states. The remaining 25 states have arcs which either name specific anticipated words or syntactic categories of words which can be accepted by that state. For example, the initial state of the network lists 13 specific words plus the lexical categories, adverb, preposition, verb, auxiliary, model, and "expletive" (including all one-word utterances such as "yes" and "okay"). Lexical categories for closed class words (articles, prepositions, pronouns, conjunctions, etc.) impose considerable syntactic restriction. Twelve states out of the 25 have only closed class category arcs or no category arcs at all. Thus these states (about half of those affecting syntactic restrictions) impose strong syntactic constraint. The remaining 13 states have category arcs for one or more of the open classes (noun, verb, adjective, adverb) and pose very little syntactic constraint for large vocabulary systems.

Mechanisms for Interfacing Syntax to Lower Levels. Current grammars derive their efficiency because of the strict control they exert. Whether they can continue to exert this strict control in parsing errorful strings is yet to be determined. Techniques for parsing in the presence of errors have not yet been developed and the ability of a parser to deliver effective syntactic support to the lower levels will depend on its ability to parse correctly in the presence of errors.

There are several acoustic cues which in turn may be helpful in the parsing of ungrammatical sentences. Pauses not only indicate word boundaries but often also indicate a grammatical phrase boundary. In addition to indicating the statement class (question, assertion, etc.), stress and intonation parameters may be able to help disambiguate when the same word may fall in several grammatical categories.

A9.4 Lexical Level

In this section we will discuss the problems raised by the size and structure of the vocabulary and various sources of knowledge that can be brought to bear to reduce the search space within the lexicon.

A9.4.1 The Large Data Base Problem. All tasks considered, except the Voice-CS, have to deal with large vocabularies. A large data base creates several new problems for speech understanding systems, the main ones being creation, maintenance, (insertion, deletion, and modification of entries), representation, and retrieval. These problems do not arise in Voice-CS because the data base is small and is already maintained within the operating system.

The creation, maintenance and the representation problems of a data base have to be faced by all data management systems. What is new for Voice-DM is that in addition to the written form of the data, the system must also maintain the spoken form of the data. There are several systems being developed for grapheme-to-phoneme transcription of data (Lee, 1968; Allen, 1970) in connection with research on reading machines for the blind. These systems can generate accurate phonemic transcription for over 90% of the data. Whether they work equally well in the transcription of proper names is not known at the present.

Another solution to the problem of representation of the verbal form is to have one or two native speakers say the utterance which is then segmented, classified, and stored in a compact form as part of the data base. The main advantage is that much of the with-in-the-word coarticulation effects would already be accounted for and need not be calculated each time. The main disadvantage is that it is likely to use 3 to 10 times more storage space than the phonemic representation. This does not mean that programs for calculating the effect of phonetic context can be eliminated altogether. They will still be needed to calculate the between-words coarticulation effects.

The problems of data representation and search mechanism are also aggravated by the requirement of speech input and output. In a keyboard oriented data management system the input is assumed to be error-free and the representation and search mechanisms are organized for the exact match situation. With voice input, even if the input utterance is syntactically correct and uses only the legal vocabulary, the representations at each level of the speech analysis are likely to be errorful and this forces a best match situation for the data base search. Minsky and Pappert (1969) show that relatively small factors of redundancy in memory size (for hash coded representation) yield very large increases in speed for serial computations requiring the discovery of an exact match in the data base. However, for the best match situation they conjecture that for large sets with long word lengths there are no practical alternatives to large searches that inspect large parts of the memory.

Thus it appears that going from typed input to voice input would require substantial increase in the search time of the data base. However, in the case of Voice-DM the prospects are not as gloomy as predicted. The strong syntactic support resulting from the highly restricted grammar eliminates the need for the search of large parts of the data base. However, if a major part of the vocabulary belongs to the same syntactic category, such as proper nouns, then there is no alternative but to search this whole vocabulary. However, we know several heuristic devices that are peculiar to speech which would reduce the magnitude of search for the best match problem.

The evidence from Vicens-Reddy system is that only 10 to 20% of the total vocabulary need to be searched in determining the best match. The rest of the vocabulary is eliminated by various heuristic devices such as gross structural comparison, vowel similarity, reordering of candidates, and so on. It seems appropriate to assume that a similar reduction can be achieved in the case of tasks considered in this report.

Estimation of Vocabulary Size. In the protocols given in Appendix 7 for Voice-CC, a total of 430 word tokens were used by the user in 10 minutes of conversations. Allowing for repetitions, a total of 165 different words were used. A plot of the cumulative number of word types as a function of the number of tokens is given in Figure A9.11. Note that the curve is definitely concave downward, as one would expect. The slope of the curve after 430 tokens is .28 or about one new word out of four uttered, so the curve is far from saturation. It is, therefore, difficult to predict the size of a stable vocabulary from this data. However, if we extrapolate this curve for two hours of conversation at the same rate (43 word tokens of user questions per minute of time) and slope, we would expect a vocabulary of about 5000 words. The figure would probably be considerably less than this because we would expect a continual decrease in slope. However, it seems clear that a stable and useful vocabulary for this task should be of the order of 2000 or 3000 words.

Given that we have an English-like vocabulary of 3000 words, it would be useful to know how these are distributed among various grammatical classes. This would indicate the expected reduction in the lexicon search given that we know the appropriate grammatical class of word to be compared from syntactic considerations. Figure A9.12 gives the percent of words in each grammatical class that were found in 2500 words of spoken vocabulary collected by Jones and Wepman (1966).

Since we are dealing with a specialized task, that of answering questions about the use of TENEX system on the PDP10, one would expect that it would be possible to subdivide further the nouns, verbs, and adjectives on the basis of semantic subtopic classes. As was illustrated in the case of Voice-CS, the effect of semantic subselection is expected to be limited, say ranging from a reduction by a factor of 1.5 to 2.

It is interesting to study the effect of increase in vocabulary on the observed phonemic ambiguity among the words of the vocabulary. This can be estimated by the use of the model described in Appendix 10. A 3200-word spoken vocabulary collected by Jones and Wepman (1966) was used in this study. A dictionary containing these words and their phonemic transcriptions was used in the computation. Out of this dictionary 4 different random vocabularies were chosen of sizes 50, 200, 800, and 2600. For each of these vocabularies, estimates of similarity score distributions were calculated by random sampling of 1000, 2000, 3000 and 4000 pairs from each of the vocabularies respectively. These percentage distributions are shown in Figure A9.13. Note that the percentage distributions in each similarity range appear to reach a stable value asymptomatically. This indicates that after the vocabulary reaches a certain critical size, any

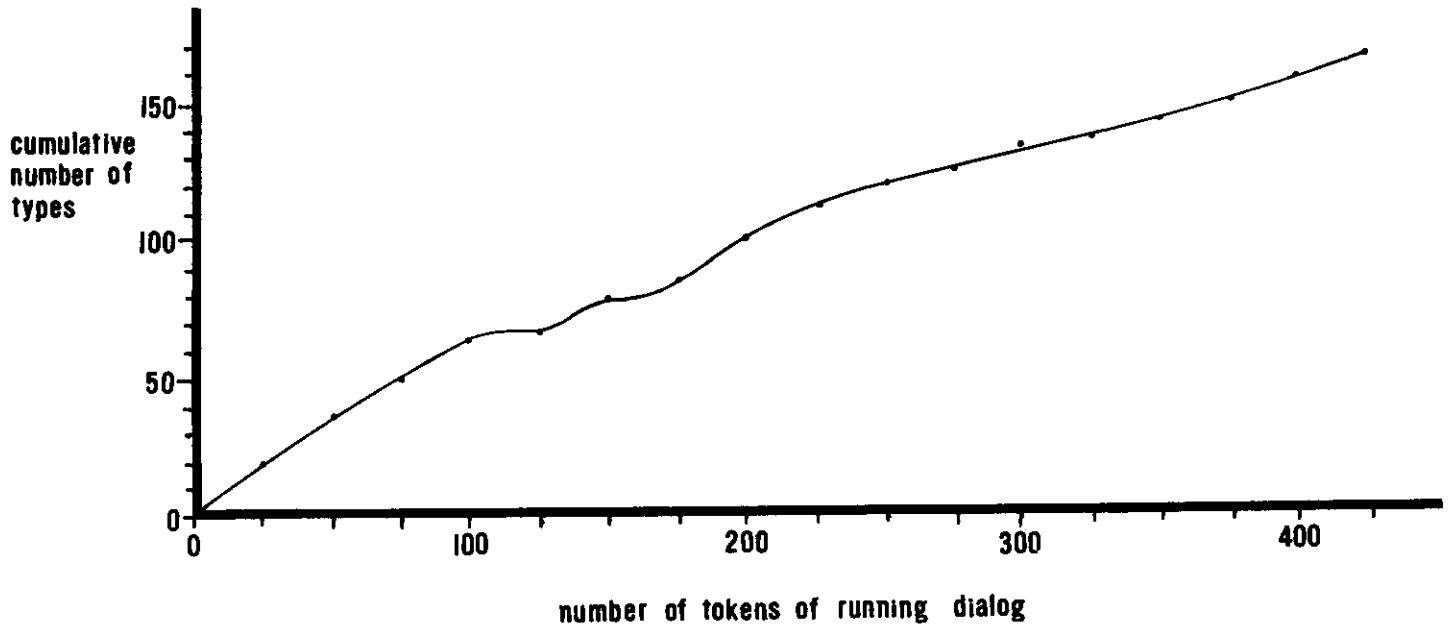


Figure A9.11 CUMULATIVE NUMBER OF WORD TYPES vs. NUMBER OF TOKENS

WORD CLASS	%
Nouns	43.7
Verbs	21.1
Adjectives	21.7
Adverbs	3.8
Auxiliaries	.9
Conjunctions	.8
Pronouns	1.3
Quantifiers	2.7
Prepositions	2.0
Articles	.1
Relatives	.6
Indefinites	.5
Interjections	.8

Number of words: 2495

Figure A9.12: Distribution by Grammatical Class.

Similarity Score Decade Range	50 Word Vocabulary (1000 Samples)	200 Word Vocabulary (2000 Samples)	800 Word Vocabulary (3000 Samples)	2600 Word Vocabulary (4000 Samples)
90 - 99	0 %	0 .%	0.1%	0.1%
80 - 89	0	0	0.1	0.1
70 - 79	0.4	0.1	0.1	0.2
60 - 69	0.8	0.4	0.3	0.4
50 - 59	0.8	0.6	0.8	0.9
40 - 49	3.2	1.8	2.6	2.7
30 - 39	8.6	5.0	5.5	5.4
20 - 29	18.7	14.6	15.3	14.5
10 - 19	35.5	29.5	33.3	31.3
0 - 9	32.0	48.3	41.9	44.6

Figure A9.13. Estimates of Similarity Score Distributions for Several Dictionary Sizes

further increases in size do not increase the incidence of phonemic ambiguity among words of the vocabulary. It was expected that after the vocabulary reaches a certain value, say a thousand or more, any further increase in size would not increase the incidence of phonemic ambiguity among the words of the vocabulary; i.e., after a certain point, the number of words in a language that are confusable with each other was expected to be constant. The simulation study showed that for vocabulary sizes of up to 3,000, the number of words that are confusable with each other appear to increase linearly. This seems to imply that one has to consider very large vocabularies before one can make any predictions about the incidence of word ambiguity within the language.

A9.4.2 The Effect of Errorful Phoneme Strings. Another important problem at the lexical level is the determination of the lexical items from a noisy phoneme string. This problem arises because the input string of phonemes has errors in it, which makes it difficult to search through the lexicon for an exact match. The correct word will be located only if all the phonemes in the string are correct. Take the simplest case: Each phoneme has an independent and equal probability p (say .9) of being correct and the word has n phonemes in it (say 6). Then the word will be correct only with probability p^n (i.e., $(.9)^6 = .53$). The probability of a word being correct falls off rapidly when it depends on all of its components being correct, even though their individual probabilities are rather high. This multiplicative relationship is worth emphasizing, since it is at the heart of the erosion of fidelity that occurs in a multi-level system that keeps converting sequences into elements. The graph of Figure A9.14 shows the curves of probability correct as a function of word length for the case of independent and identical probabilities of correctness. Although this can be complicated in various ways (non-independent, non-equal errors) this case shows effectively the fall off of accuracy that occurs in any system that demands simultaneous accuracy of a set of components.

The antidote for this is a source of knowledge that allows the errors to be corrected. The dictionary provides this by establishing a finite set of possible strings as the absolutely correct ones, inviting us to consider phoneme strings which deviate from the true strings to be in error and to indicate that, in fact, the true string that was uttered* was one close to the received string. The notion of closeness can vary, but is determined through some kind of matching procedure between the input phoneme string and the candidate phoneme string from the dictionary.

If there were a word in the dictionary for each possible phoneme string, then, of course, no correction would be possible. Thus, the amount of error correction delivered by a dictionary runs inversely as the size of the vocabulary: the smaller the dictionary the more error correction. Most of the mechanisms associated with the lexical level serve to reduce the effective dictionary size. In particular, the selection of a subgrammar according to the estimated user state works by selecting out a subvocabulary.

There are two effects working jointly here that must be distinguished. The first, under discussion above, is generating of more specific knowledge by the restriction of the set of words that are candidates for a given phoneme string. In parallel with this is an efficiency issue, for the more vocabulary items that must be considered, the more processing is required. By eliminating candidates that would fail in any event, the efficiency can be improved. The selection of a subpart of the subgrammar according to reliable features of the input string is a mechanism of this kind -- devoted to improving the efficiency without decreasing the error (in fact, adding slightly to the error, since the selected subpart will occasionally not contain the true candidate).

A9.4.3 Sources of Knowledge. A simulation was performed (using the model given in Appendix 10) to estimate the effect of being able to select various subvocabularies of the Voice-CS lexicon. This was done by selecting a subvocabulary, determining the confusion matrix among the words for that subvocabulary, and summarizing the distribution of scores. Figure A9.15 shows the distributions of confusions (similarity scores) for the total vocabulary and for each of the subsets of vocabulary in the semantically limited grammars. If we assume that the probabilities of error are the same as those determined by the model in Appendix 10, then it becomes possible to compute the probability of error resulting from the distribution of similarities, which range from .16 to .29. This probability of error is also shown in Figure A9.15. This calculation ignores the fact that the words would actually occur at conversation with different frequencies and are not equally likely. In addition to the semantic subselection of vocabularies, we also have another source of knowledge--that provided by reliable features of the phoneme string. This, of course, does not employ the error rates since it lets through (by design) precisely those candidates that will get high similarity scores. The effect of this phonemic subselection

---

* Or was intended to be uttered, for errors can arise within the speaker as well as in the lower levels of the recognition system.

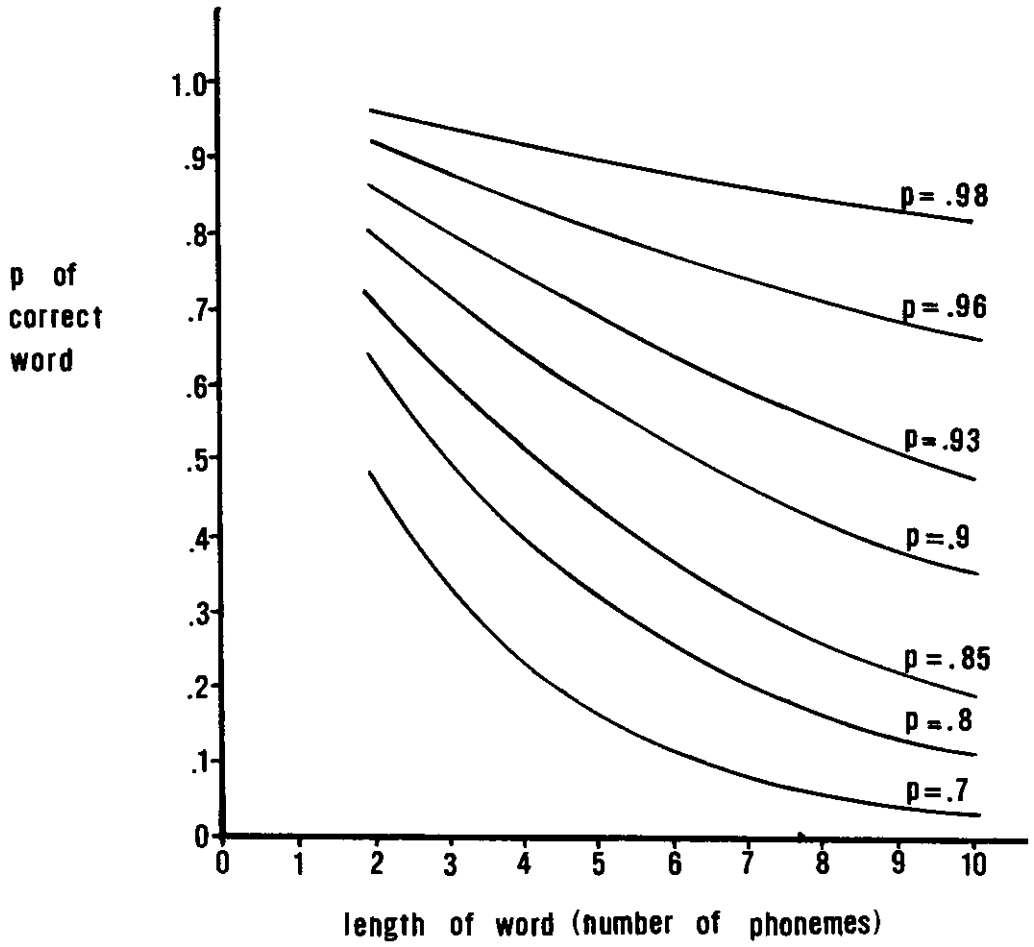


Figure A 9.14 Curves of Probability of correct word recognition as a function of word length (in phoneme) for the case of independent and identical probabilities of correctness for individual phonemes in the word.

Decade Range	Total Voice-CS vocabulary 68 words	Request-system subvocabulary 49 words	Request-job subvocabulary 47 words	Request-resources subvocabulary 38 words
90 - 99	4	2	3	2
80 - 89	4	3	0	1
70 - 79	3	2	2	3
60 - 69	6	6	1	3
50 - 59	17	10	8	8
40 - 49	46	33	26	26
30 - 39	90	63	46	51
20 - 29	191	127	100	93
10 - 19	471	257	249	199
0 - 9	1446	673	646	317
Total	2278 $= \frac{68 \times 67}{2}$	1176 $= \frac{49 \times 48}{2}$	1081 $= \frac{47 \times 46}{2}$	703 $= \frac{38 \times 37}{2}$
Total Probability of Error Using Probabilities from Figure A10.5 (combined Error	.315	.293	.166	.233

Figure A9.15. Distributions of Similarity Scores and an Estimated Total Probability of Error for the Voice-CS Vocabulary and its Three Subvocabularies

is to limit the amount of effort required by the match routine. We can estimate this with the same model simply by modifying the match slightly and cutting at a value known to include most errors. We get a figure of about 60% reduction, which compares with a figure of about 80% for a similar mechanism on the Vicens-Reddy program.

A9.5 Phonemic Level

Identification of the phonemic string is subject to errors just as in the lexical level. In this section we will consider the problems raised at the phenomic level and the sources of error and knowledge present in solving these problems.

A9.5.1 The Lexical Segmentation Problem. The string of phonemes that arrives at the lexical level to be processed consists of an unending sequence of phonemes without any marking for the boundaries between words. This problem has been considered to be quite serious (as have all problems of segmentation, at whatever level). Little is available in the literature to help with an objective assessment of the problem. For instance, although the Vicen-Reddy program, as described in Section 4, ostensibly did some segmentation, it was of a very rudimentary kind that cannot be extended.

Our description of an unending sequence of phonemes overstates the problem. Words do run together, but pauses of substantial duration also occur that make it quite clear that a word boundary occurs. (Short periods of silence are of no help whatever, since these may be associated with the utterance of stop consonants.) There is reasonably good evidence on the pause structure of human conversation (Goldman-Eisler, 1968). The mean number of words between pauses ranges from 3 to 12, being affected by the nature of the speech (whether it is spontaneous or learned) and by the individual (strong consistencies with individual). For instance, when describing or interpreting cartoons, 50% of speech occurs in phrases of 3 words or less, 75% of phrases of 5 words or less and 90% of phrases 10 words or less. Thus, the pauses do about 20% of the job of segmenting. More important, however, they provide error free bounds within which the other segmentation process can work.

Within a phrase words run together without any separation at all. That is, a phrase, such as "how are you?" is spoken in the same manner as the three syllable word "Waterloo." This is especially true of overlearned phrases, which may come to function as extended words. But it is also true in general. If someone asks "Where did Bill go?" the answer "Bill went home" is likely to be said as a single continuous utterance. Figure A9.16 shows the speech waveform of "How are you," from which it is clear that word boundaries in the form of diminished energy (silence-like events) simply do not exist.

Let us consider first the simplest case, namely a phrase without silence markers, but otherwise phonemically correct. E.g., our two examples above would appear (using the standard phonetic alphabet) as:

/ H AU AA R Y OO /

/ B I L W E N T H O M /

The obvious decoding strategy is to attempt to segment the phonemic phrase by means of the dictionary. That is, match the first part of the string to the dictionary, then segment where the dictionary word says, then match the new first part to the dictionary again, and so on. One will be led down some blind alleys, but hopefully, these would eventually lead to nonsense (i.e., no word in the dictionary for the supposed next word.) This strategy has been evaluated (Reddy and Robinson, 1968) and the expectations are satisfied. The correct sequence was determined every time and without excessive computation. (Unfortunately for our purposes, no detailed performance data are given; e.g., the amount of back-up that occurred is unknown.)

Two things are wrong with this simple case: generally, the phoneme string is not error free; and particularly for Voice-CC and Voice-CS, not all the words encountered in the phrases are in the dictionary. This latter issue will be discussed in the next section; here we will concentrate on the problem of error.

We can think of the problem in the following way. At some point we have segmented the phoneme sequence successfully (i.e., all segments selected have matched successfully to some word in the dictionary), even though (as we shall assume) the segmentation is actually wrong. We now need to consider all possible next candidates for segmentation. Figure A9.17 shows the situation. There is one candidate that is one phoneme long, another that is two, another three and so on. For each of these candidates, there is a probability that it matches a word in the dictionary, even though (per assumption) the proposed segment does not correspond to an uttered word. In the error free case studied by Reddy and Robinson, this probability is only that of a subsequence of an utterance forming another legal word. In a more realistic case, errors in phoneme identification and variation in the representation increase that probability. If, due to the error variation, we actually get a match, then the picture of Figure A9.17 repeats itself at the next segment. If we don't get a match for any of the extensions, then we know that this branch is not possible and we can back down to the prior situation. Thus, the

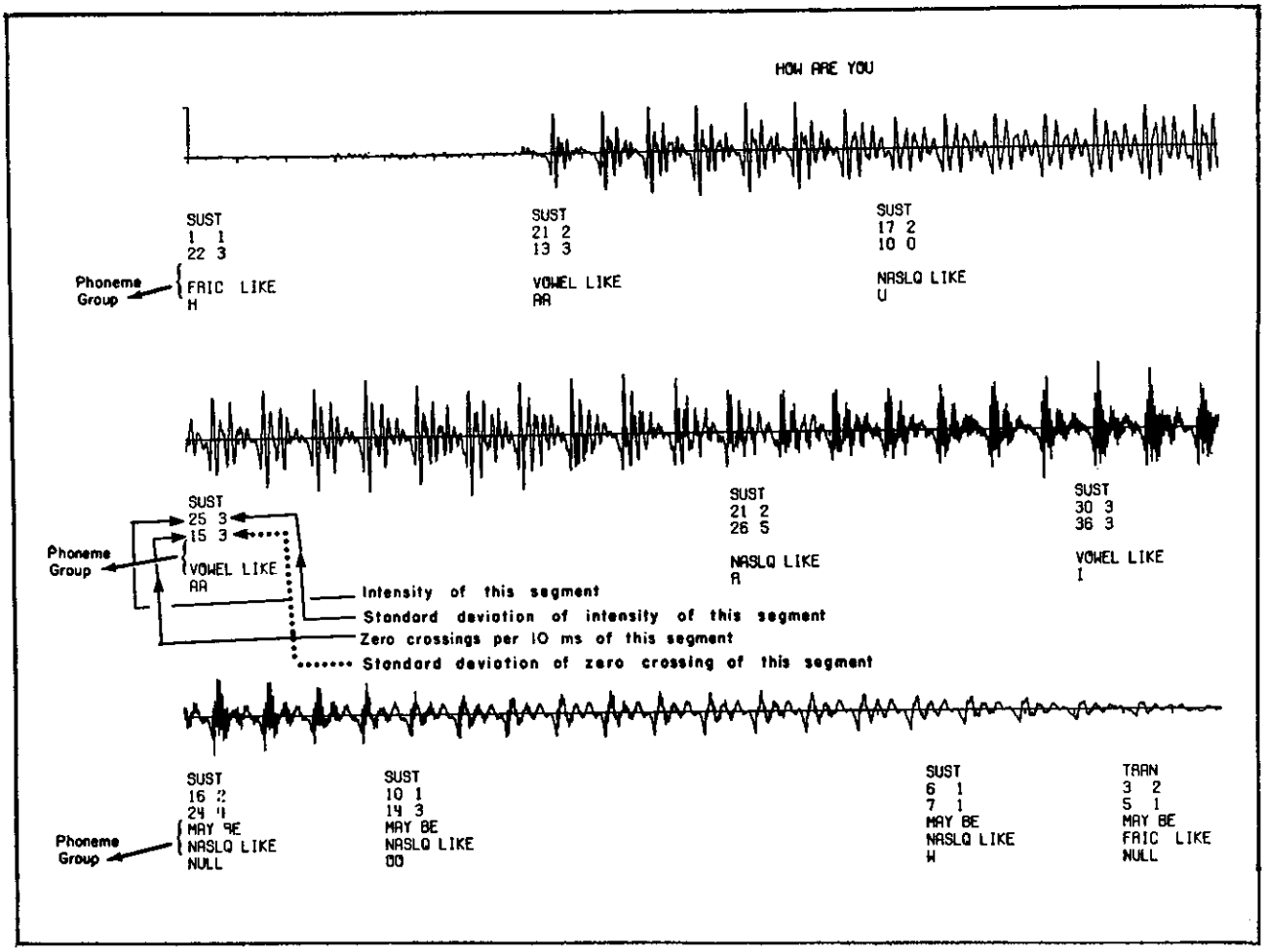
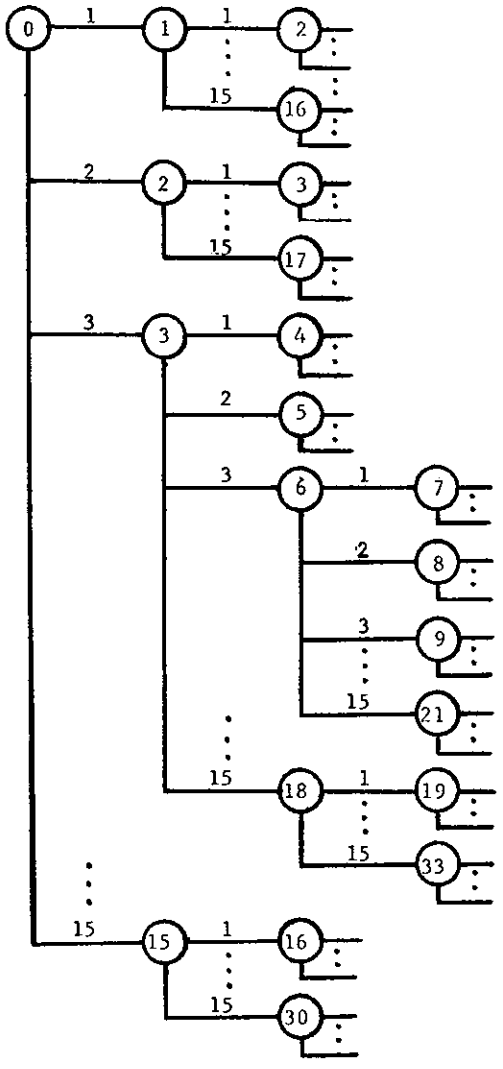
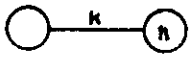


Figure A9.16: The speech waveform of "How are you" shows that word boundaries in the form of diminished energy (silence-like events) simply do not exist.





k = number of phonemes in next segment

(n) = total number of phonemes covered

k = 15 is an arbitrary upper bound. Theoretically one could try all the phonemes in the sequence as a single lexical item.

Figure A 9.17 Tree of Segmentations of Continuous Speech.

situation forms an expanding tree of possibilities. Each node of the tree is essentially identical, until we get down toward the end of the total string, where there are not enough phonemes left for the long candidates. Thus at the end there are fewer candidates for extension than during the major part of the segmentation.

The key question is the expected number of branches that will survive at each node (assuming the node to be a false one). To evaluate this we need to know the probability of a false match, given that we take a subsequence of phonemes out of the middle of a continuous string and match it against the Voice-CS dictionary. Figure A9.18 gives some estimates of these probabilities. These were computed from our model by putting together continuous phoneme sequences from concatenations of the phoneme sequences of words from a base dictionary. To model the fact the Voice-CS permits many words in the input stream that are not in its dictionary, this base dictionary consists of a particular set of 3000 words which most frequently occurred in a sample of normal (monologue) speech (Jones and Wepman, 1968). Substrings were selected at random for this continuous sequence of phonemes and were matched against the Voice-CS dictionary. Since the words were not in isolation, the bounding phonemes (which were silences, in the isolated word cases discussed earlier) were taken to be the bounding phonemes in the running sequence. Figure A9.19 shows the technique.

The last line of Figure A9.18 contains the probabilities of error, derived from the Class B data of Figure A10.5. These estimates are derived from data on runs of the Vicens-Reddy program made after first training the system on several (3-8) other speakers. It is felt that this is the best data to use; Class A represents a single speaker and is "too easy" while Class C represents training of the system and is thus "too hard." As can be seen from Figure A9.18 the probability of error is about .3 for the single phoneme case and then quickly falls off toward zero as the number of phonemes gets about 6. Thus, the expected number of continuations from a node (given that the segmentation is, in fact, incorrect) is the sum of the probabilities of error at each length which is about .85.

Now the combinatorics are with us, since a sequence of .85 probability events must occur in order to obtain a total false decomposition. The number of such events ranges from 2 (for a single word bounded by silences) up to about 100 (for an attempt to decompose a 10 second utterance into many small words). Sequences of longer than about 10 words can be safely disregarded. Sequences of less than this have some chance of producing error. But for these the end effect must be taken into account, since the probabilities of success are smaller than .85, due to the truncation of the sequence.

The above calculations point out two things. First, that the knowledge that the whole sequence must consist of words provides substantial constraint in segmenting speech. Second, this help falls into three classes. If the expected number of false continuations is small enough (say .4 instead of .8) then this mechanism can do most of the work by itself, for the sequences squeeze off very fast. If, on the other hand, the expected number is greater than 1, then the number of false sequences grows very rapidly and there is real trouble. There is a small region in between (in reality, where our simulated numbers actually put us) such that slight improvements in the matching and handling of other errors in the system make substantial differences in the performance on segmented speech.

The above calculation is also a base calculation in not expecting any direct help in terms of local clues for segmentation (or semantic clues, for that matter, since we applied no constraint to the string other than that it be a sequence of legal lexical items). There are in fact some such clues, though there has been little systematic study of them for recognition purposes.

The main sources of knowledge on local clues for segmentation are that (1) certain phonemic sequences cannot occur within a word (Siversten, 1961), (2) suprasegmental features, such as duration, pitch, and amplitude, exhibit different characteristics if there is a word boundary between two segments than if there is not (Lehiste, 1970) and (3) coarticulation effects across word boundaries are much less dominant than within a word (Lehiste, 1964). The main difficulty with these sources of knowledge is that they are in generative form and their analytic counterparts appear to be much harder to formulate.

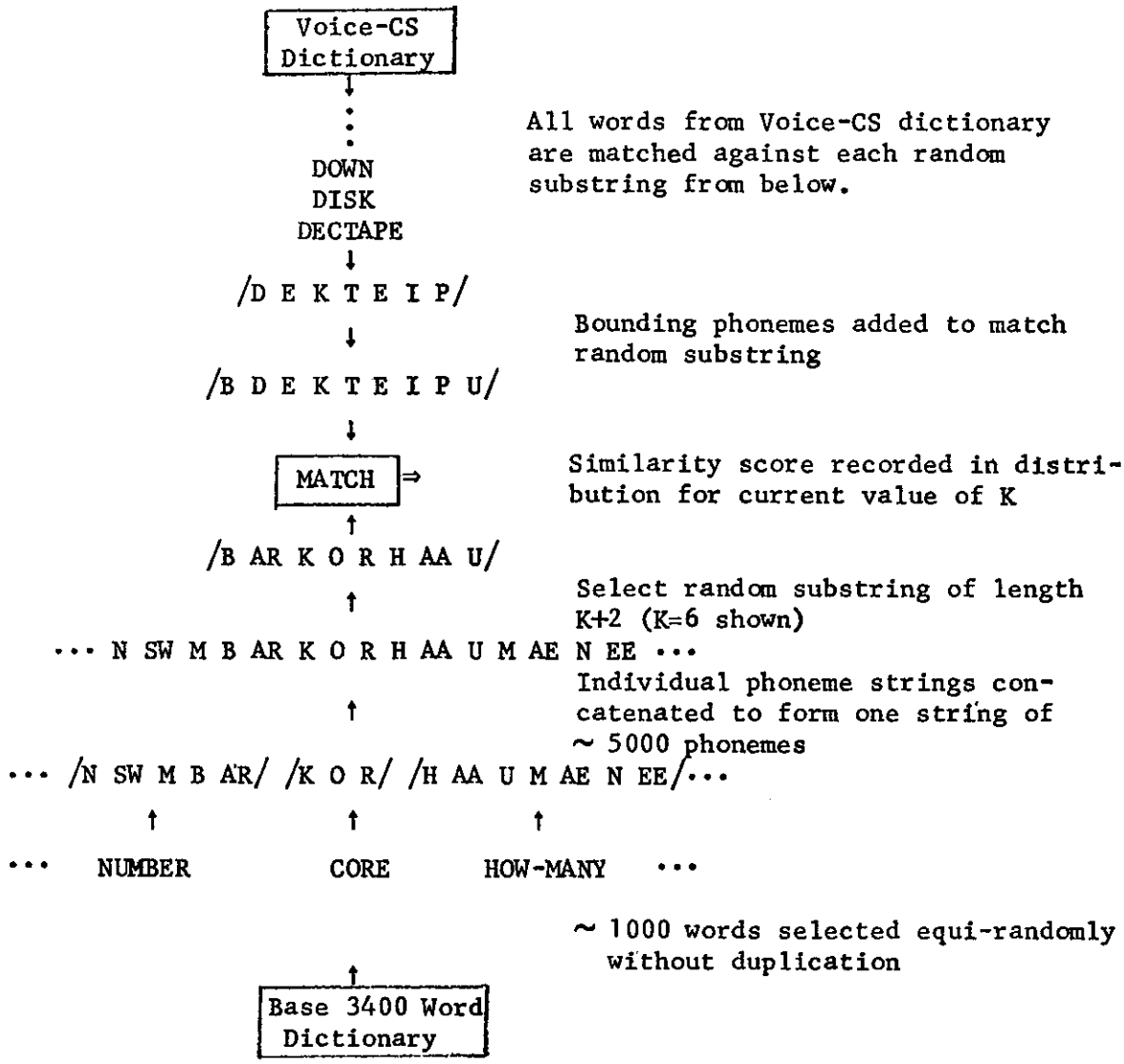
An important consideration, not apparent in the discussion so far, is that English has a syllabic structure. This shows up as a somewhat periodic vowel - non-vowel alternation at the phonemic level. Thus, in considering the possible matches in the combinatorial attempt to decompose a continuous phoneme string, the true length is more like the number of syllables than the number of phonemes. The numbers if Figure A9.18 take this into account somewhat, since they are averaged over all cases, but this reduction in combinatorial complexity should be noted.

A9.5.2 Errors in Phonemic Strings and the Multiple Labels Problem. The probabilities of error at the phonemic level (addition or deletion of phonemes from the string or incorrect recognition of a phoneme in the

Decade Range	Random Substring Length -- k										$\Sigma(P(E_k))$
	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	
90-99	6	0	2	0	0	0	0	0	0	0	
80-89	5	3	1	1	0	0	0	0	0	0	
70-79	5	6	3	0	0	0	0	0	0	0	
60-69	17	18	3	0	0	0	0	0	0	0	
50-59	33	38	7	9	0	0	0	0	0	0	
40-49	57	119	46	11	2	1	1	0	0	0	
30-39	138	206	137	54	3	1	2	0	0	0	
20-29	185	365	359	160	50	11	1	0	0	0	
10-19	327	635	759	626	231	117	19	5	2	0	
0- 9	2627	2010	2083	2539	3114	3270	3377	3395	3398	3400	
Probability of error -- Class B (derived from Fig. A10.5	.307	.252	.168	.104	.009	.003	.001	.0004	.0002	.0000	.85

For each value of the random substring length (k), 50 different random substrings of the pseudo-continuous phoneme string were each matched against all 68 words in the Voice-CS vocabulary, giving 50x68=3400 similarity values in each distribution.

Figure A9.18. **Distributions for Simulation of Continuous Speech Segmentation and Estimates of the Probability of a False Match**



Notes: For a single value of K, ~ 50 different random substrings are selected and compared with all words from the Voice-CS dictionary.
The process is iterated for K=1,2,...,10 and a separate distribution output for each value of K.

Figure A9.19. Technique for Segmentation Simulation

string) has led many researchers to attempt to avoid intermediate recognition of the phonemic level by defining lexical items directly in terms of a parametric representation. This tactic precludes using the sources of knowledge available about the phonemic structure of speech and it assumes (perhaps falsely) that the parametric representation of a word does not change radically depending on sentence environment. The main source is the finite alphabet of phonemes, equivalent to the word dictionary at the lexical level. From one point of view the situation is much better at the phonemic level: There are only some 40 phonemes (though if we add the suprasegmental indicators, such as stress and intonation, the number creeps up by some modest, but unknown amount). This is to be compared with a lexical dictionary of, ultimately, thousands of entries. However, from a second viewpoint the situation is worse, since the relation of the phoneme to the parametric level is substantially more complex and less well understood. Furthermore, what is understood is derived primarily from analyzing the system that produces speech. The human recognition system is not even identified psysiologically beyond the peripheral apparatus that converts sound into neural impulses. There is knowledge of recognition that comes from behavioral experiments. Most of these do not relate directly to the acoustic representation of speech, as it must be dealt with by a mechanized recognition system. Some work, however, has been done with synthesized acoustic signals (Flanagan, 1965: see chapter 6 for a summary of the known results on synthesis and chapter 7 for the results on perception) and these experiments form a major source of what information we do have on the recognition-significance (to humans) of the speech signal.

A few attempts have been made to recognize phonemes by machine (Reddy, 1967; Medress, 1969; Tappert, Dixon, Beetle, and Chapman, 1970). They are not all at the same level of sophistication nor use the same techniques (though all make no use of higher constraints). Success rates vary from below 50% to around 90% depending on many factors. The lower scores come from tests on new data from different speakers for a full phonemic alphabet of 40 symbols; the higher scores come from successive restrictions to generality in various ways.

Let us consider the situation in more detail. Figure A9.20 shows a fragment of a protocol for the Voice-CS task with a set of phoneme identifications. These were generated by hand according to a scheme developed by the Forgies (Forgie and Forgie, 1962). The total acoustic stream is segmented and for each segment a list of phonemes is given along with a confidence score (ranging from 0 to 100 independently for each possibility). The entire phrase coded was:

> "Ha ha.. Ok. um. (b) let's see.. Wh What things would I want to know right off hand.. I'd like to know how many users are on the line? On.. are on the system. Thats obvious."

The total encoding is given in Appendix 11.

Without yet examining causes, it is apparent that there is considerable uncertainty about the exact identification. The first choice does not often lie far above its competitors in confidence. In fact, if we compute the simple error score of first choice against the correct phoneme (from the transcription column) we get 55% correct (49 out of a total of 89 phonemes), a figure comparable to the 50% - 90% figures just quoted above as representing the state of the art. (For instance, the noise quality is poor, since it is taken over a regular telephone line.)

Combinatorial Explosion Problem of Multiple Labeling of Segments. Keeping the total set of plausible identifications preserves a substantial amount of information about the utterance. However, it is useful only if that information can be extracted. For instance, if one simply generated all the candidate sequences in order of composite plausibility, the combinatorics would be prohibitive (e.g., 5^{10} if there are 5 plausible identifications for each of 10 phonemes).

One way to handle such data is to carry forward the entire ensemble until the information appropriate to using it is available. This turns out to be at the lexical match, where, instead of matching against a single given phonemic string, the match should be made against the entire array as given in Figure A9.20.

Another way to reduce the combinatorics is to use the first choice for each segment to compare against the whole vocabulary of Voice-CS and then choose only those words with high enough scores for comparison with all the other combinations. Figure A9.21 shows the flow chart of this technique for the specific example of "users."

Figure A9.22 gives the results obtained by the use of the model. The words "busy, date, is, k, minutes, one, six, space, state, tape, users, what, where" were chosen as the most likely candidates. It can be seen that the word "users" was the first choice in every case. The word "space" was the second choice 95 times and the word "minutes" was the second choice 74 times.

Transcription	Segment Code	Phoneme Scores	English
AI	V	UH 60, E 56, U 56, AH 52	I'd
D	VS		
L	N	L 70, M 50, R 50, N 40, W 40	like
AI	V	AI 88, AE 70, E 74, UH 67, AH 64	
	S		
K	PL	K 50, T 47, P 42, D 26, G 20, B 21	
	S		
T	PL	T 70, K 64, P 40	
U*	h	This appears to be an unvoiced vowel for which we have no analysis.	to
	S		
N	N	N 70, NG 45, L 40, M 36	know
O	V	OU 68, O 64, AW 56, U 56	
H	VF	DH 50, V 40, H 40, ZH 34, Z 30	how
OU	V	OU 70, O 60, AW 58, U 52	
M	N	M 60, N 56, NG 45	many
E	V	AI 76, AE 65, E 62, UH 55	
N	N	N 70, NG 60, M 48	
EE	V	AI 72, EE 69, AY 65, I 60	
Y	SU	Y 70, L 50, W 30 Good chance of missing this segment	users
OO	V	U 60, UH 58, I 50	
Z	VF	Z 80, S 70, ZH 68, SH 60, DH 40, V 40	
ER	V	E 70, I 67, UH 55, ER 52	
Z	FR	S 80, Z 70, SH 68, ZH 60, TH 40, F 40	
ER	V	ER 74, AW 74, AE 71, AH 62	are
AH	V	AH 88, UH 72, ER 66, AW 50	
N	N	N 70, M 59, NG 54	on

Figure A9.20. Phonemic analysis of Voice-CS Protocol Fragment

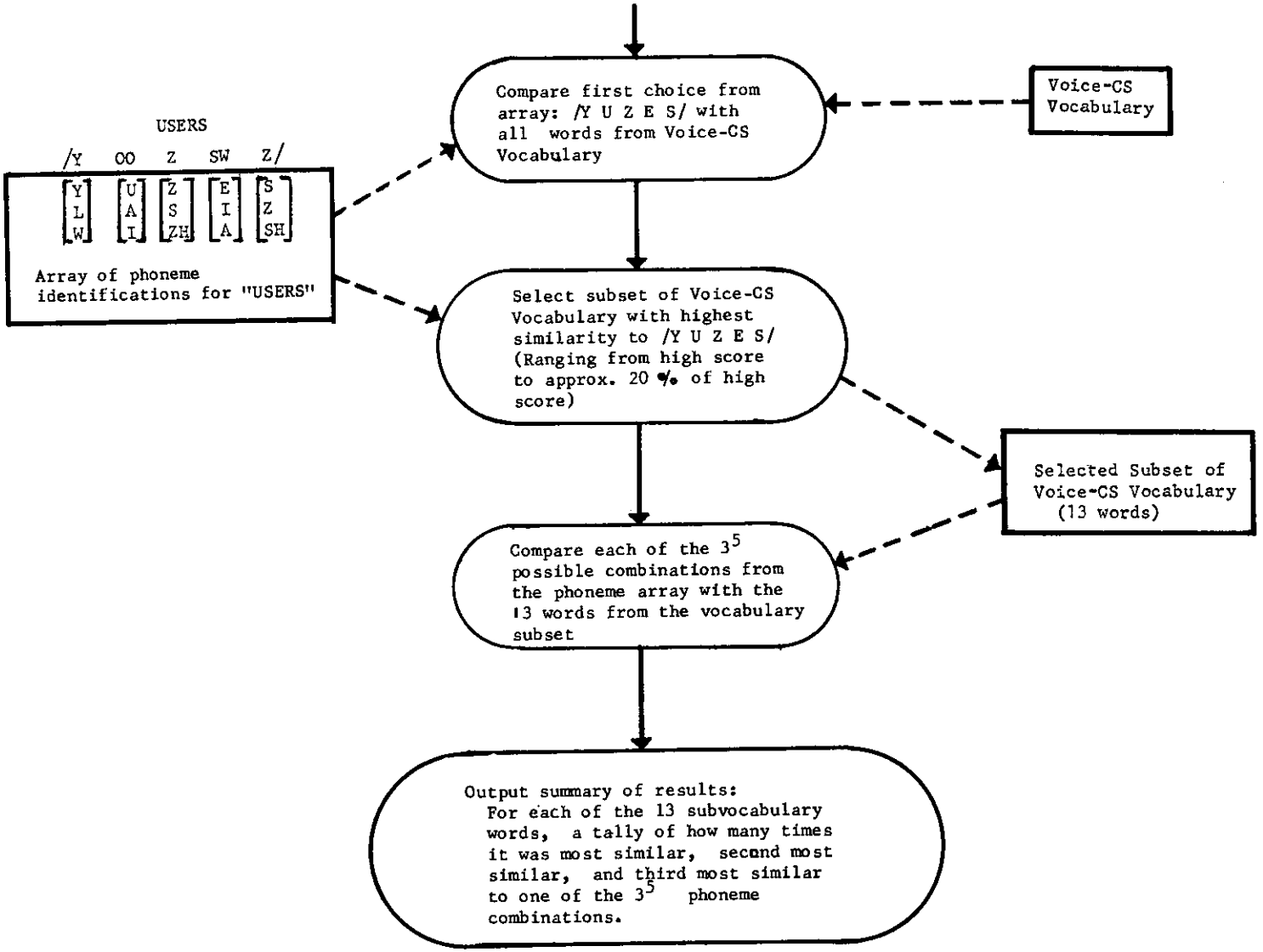


Figure A 9.21: Technique for Matching Phoneme Array for "USERS" Against Voice-CS Vocabulary

Voice-CS Selected Subvocabulary	No. of times word was most similar to one of the "USERS" phoneme combinations	No. of times word was second most similar	No. of times word was third most similar
BUSY	0	24	61
DATE	0	11	11
IS	0	2	9
K	0	0	0
MINUTES	0	74	36
ONE	0	37	47
SIX	0	0	0
SPACE	0	95	49
STATE	0	0	26
TAPE	0	0	0
USERS	243	0	0
WHAT	0	0	4
WHERE	0	0	0

The 13 word subvocabulary shown was selected as the words from the Voice-CS vocabulary with the highest similarity to "USERS".

Figure A9.22. Results of Matching Phoneme Array for "USERS" Against Selected Voice-CS Subvocabulary

A9.6 Parametric Level

A parametric representation of the speech waveform is an information-reducing transformation from the speech signal into a small set of waveform characterizing functions. The parametric representation should bear a straightforward relationship to speech production constraints in order to facilitate recognition of phonetic features and to decode the phonological rules of speech production.

For example, a representation that satisfies these properties characterizes the speech waveform in terms of the 4 or 5 resonant frequencies of the vocal tract (Formants), resonance bandwidths (or amplitudes in the case of frication excitation), the amplitude of voicing, the amplitude of aspiration, the amplitude of frication, and the fundamental frequency of voicing. Automatic extraction of these parameters at a low error rate is not yet state-of-the-art, but experiemental systems are approaching reasonable performance levels (Schafer and Rabiner, 1970).

Other parametric representations such as the axis-crossing frequencies and amplitudes of Reddy (1967), rectified, smoothed, and quantized filter bank outputs (e.g., Stevens and Bismark, 1967), and the ASCON parameters of Culler (1969) are not as easy to relate to speech production. In addition, axis-crossing and filter bank parameters may not contain sufficient information to implement some algorithms successfully.

Given that an adequate representation of speech at the parametric level has been chosen, it becomes possible to discuss the problems associated with the parametric level in terms of this representation. Here we will discuss three main sources of parametric variability: (1) variability resulting from segmental context, (2) variability resulting from sentential context (stress and intonation), and (3) parametric variability resulting from speaker characteristics.

A9.6.1 Parametric Variability Resulting from Segmental Context. One of the major sources of error affecting the representation of the utterance at the phonemic level is the variability of segmental parameters of a given phoneme in different contexts. Various acoustic realizations of a phoneme (allophones) can exhibit radically differing characteristics depending on context. Consider some of the alliphones of the phoneme /T/* that appear in words of the Voice-GS vocabulary: time, status, printer. Normally /T/ is manifested in terms of a silence segment, followed by an aspiration, followed by a transition into the following vowel as in the word time. However, any one of the three cues may be missing. The aspiration segment might be missing, e.g., the /T/ following /S/ in status. The silence segment might be missing, e.g., the /T/ following /N/ in printer. The silence segment might be voiced, e.g., the intervocalic /T/ in the word status. Some of the transitional cues might be missing if there is a word boundary, i.e., if the /T/ is in the word initial or word-final position. To consider another example, the characteristics of /OO/ in the word two are usually very different from the expected characteristics of /OO/ during the first 100 milliseconds from the onset of the vowel. Thus it would be impossible to classify the vowel segment as /OO/ without taking into account the effect of coarticulation.

Parametric variability resulting from phonetic context is usually explained by considering the complex articulatory gesture that results from the given sequence of phonemes. In general, two articulatory gestures corresponding to two adjacent phonemes overlap in time. This overlapping articulation of adjacent phonemes is called coarticulation. At any given instant the observed segmental parameters are the direct result of a coarticulation of the different gestures.

There have been intensive attempts to predict the effect of coarticulation by means of acoustic-phonetic rules (Lindblom, 1963; Öhman, 1968; Stevens, House, and Paul, 1966; Broad and Fertig, 1970). These rules are usually in a form suitable for the generation of speech, rather than for analyzing incoming speech. This has led Stevens and Halle (1962, 1964) to suggest "analysis by synthesis" as a model for speech recognition. This model for speech recognition involves a comparison of the input spectrum with some internally generated spectra, and an error signal fed back to the generator for the next stage of analysis-by-synthesis.

If most of these generative rules can also be expressed in an analytic form, then the computationally more economical "hypothesize-and-test" might be more suitable. This technique involves hypothesizing the presence of a phonemic sequence and formulating or selecting a test that would verify the hypothesis. This is one of the methods that has been used successfully in artificial intelligence literature (Newell, 1969). In the extreme, that test could be equivalent to the comparison of spectra in analysis-by-synthesis with no reduction in the computational effort, but usually this is not the case; e.g., it is not

---

* We use computer-style phonemic notation for consistency with earlier sections.

necessary to generate the whole formant trajectory when a simpler test of the slope can provide the same information.

Whether the acoustic phonetic rules can be adequately expressed in analytic form is at present open to question. Consider, for example, the following rule which at first glance would appear to be in a form useful for recognition:

> If a nasal-like segment is followed by a burst-like segment, then the burst could represent a stop phoneme even if there is no intervening silence segment; further, the nasal and the stop will have the same place of articulation.

Considering a specific example, this rule could be expected to predict the expected cues of the sequence /NT/ in printer. However, similar cues would have also resulted from the phonemic sequence /NS/.* Thus, not only the rules but also their exceptions and the ordering among the rules become important factors in determining their applicability and effectiveness.

At present there exists no systematic codification of acoustic-phonetic rules that can be used either for generation or analysis. Many of the rules exist only in the heads of researchers in the fields. Many others are yet to be discovered and there has been no systematic attempt to codify all the known knowledge about the acoustic phonetic rules in the form of a book.

In limited language situations, such as Voice-CS, such a book of general acoustic-phonetic rules, even if it existed, would be of limited value. One would not want to program-in all the rules in the book when only a few of them are relevant. Furthermore there will exist other rules which are specific to this limited language which cannot be generalized to all of English and therefore would not exist in a general book. It appears possible in theory that an automatic system capable of generating acoustic-phonetic rules of a limited language can be programmed on a computer. Some of these may also be extracted from a set of kernel utterances by means of "analysis-by-learning" techniques.

"Analysis-by-Learning" is one of the methods that has been successfully used in artificial intelligence research. It involves abstraction of useful information from several examplars. Thus if the phonetic realization of a given sequence of phonemes is not known as a theory, then the computer attempts to extract the appropriate tests by examining the parameters of several utterances containing that phonemic sequence. The overall structure of the test would be preprogrammed from the known linguistic knowledge, and the specific details of the test would be filled in by the computer from the examination of the data. No assumption is made that the test so derived is complete or predicts the behavior for all realizations of that phonemic sequence. All that can be said is that if the test is satisfied then it is very likely that it is a result of that phonemic sequence.

There is also a problem of validation of acoustic phonetic rules. That a new rule has been proposed and tested for acceptability by means of a synthesis experiment does not imply that all (or even most) speakers will exhibit acoustic realizations as predicted by the rule. Such rules are usually sufficient conditions but not necessaty conditions. The observed differences in the characteristics of stop consonants in a synthesis experiment (Delattre, et al 1955) and in an analysis experiment (Halle, Hughes, and Radley, 1957) illustrate the point.

A9.6.2 Parametric Variability Due to Syntactic and Semantic Context. Segmental parameters of a phoneme are affected not only by the phonetic context but also by morphemic, syntactic, and semantic context of the utterance. Acoustic characteristics of the same word (and thereby the phonemes in the word) can exhibit radical differences depending on the sequence in which it appears. Figures A9.23 and A9.24 illustrate how some segmental features appearing in words spoken in isolation do not appear in the same words spoken in sentences.

Figure A9.23 illustrates how the syllabic /L/ in the word "decimal" is reduced to a schwa in the semantic context of "The number is decimal two hundred and twenty-two." Figure A9.24 provides another example of vowel reduction in the word "divide" in different contexts.

Most of this behavior is rule-governed and, to that extent, can be deciphered from a knowledge of English Phonology.* For example, the noun pluralization morpheme, /S/, is realized by the phonetic segments /SW Z/, /S/, or /Z/

---

* There may be other cues, such as duration, which would help in disambiguation in this situation. However, such additional rules can only be activated after the realization of the possible ambiguity that could result from this rule.

* Phonology is a science dealing with the history and theory of sound change in a language--in particular, sound change resulting from morphemic, syntactic and semantic context of the sentence.

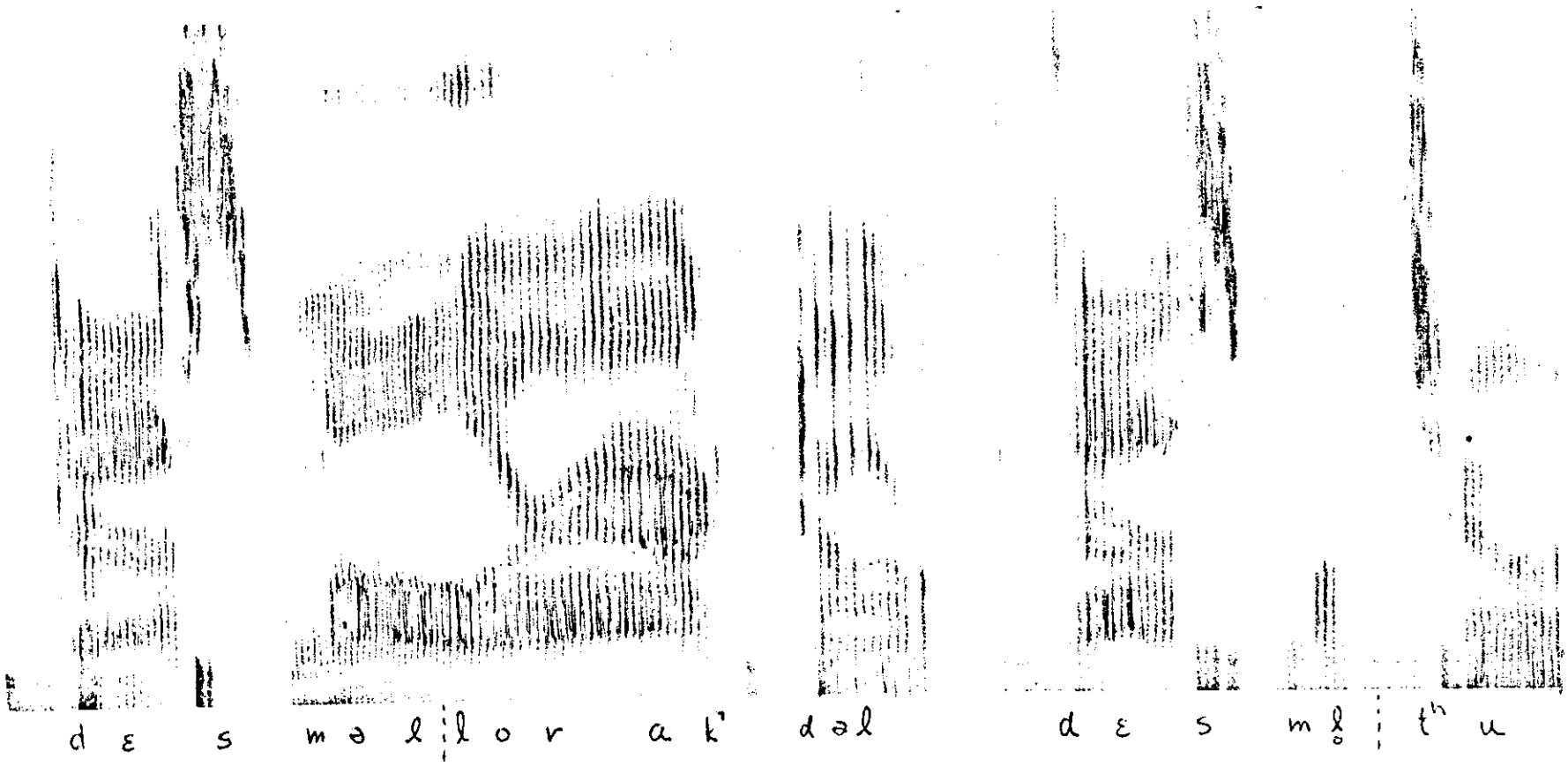


Figure A9.23: Sound spectrograms excerpted from the two sentences: "Decimal or octal?" and "The number is decimal two hundred and twenty-two." The same speaker uttered the sentences in a normal clear speaking voice. The differences between the two versions of "decimal" are due to the sentence context in which they appear.

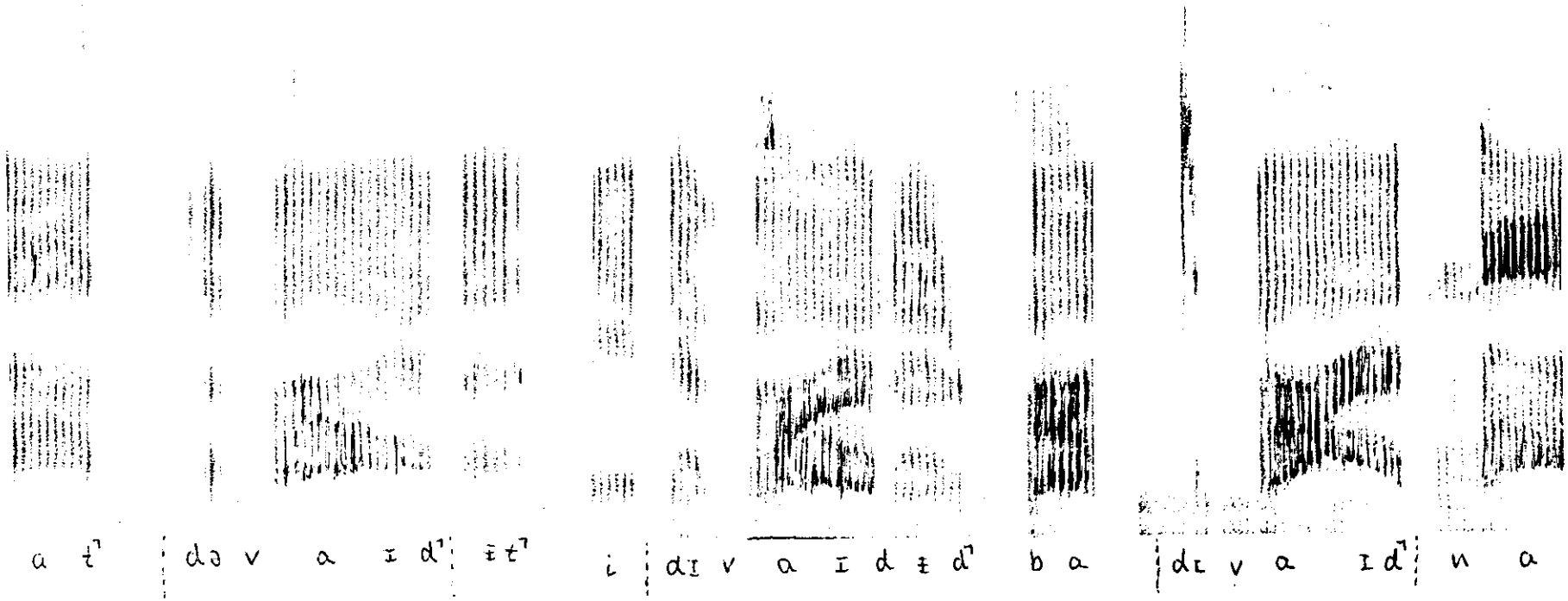


Figure A9.24: Sound spectrograms excerpted from the sentence: "He should multiply, not divide it.", "Thirty three divided by seven," and "Divide nine by seven." The same speaker uttered the sentences in a normal clear speaking voice. The differences in the word "divide" are due to the sentence context in which it appears.

depending on whether the final consonant of the noun has the feature + strident (fish+/SW Z/), or is -voiced (cuff+/S/), or is neither (bear+/Z/). In the final case, the /Z/ will be devoiced if it is followed by a voiceless consonant or a pause. In this case, the cue for decoding the fact that a /Z/ was uttered, rather than /S/, is that the duration of the vowel nucleus preceding the /Z/ is lengthened over its typical duration by about 50%.

Other phonological rules give the sentence its rhythm and stress pattern. For example, phrase and clause boundaries are often delimited by pauses and preceded by slowly falling average fundamental frequency contours (Lieberman,1967; Lehiste, 1970). Strongly stressed syllables are of relatively longer duration and have fundamental frequency contours that start at a high frequency and fall rapidly at the end of the syllable or during the next syllable (Cushing, 1969). The success of Voice-CS depends to a great extent on the hypothesis that semantically meaningful keywords are usually also the stressed words in the utterance, and the fact that stressed words are least susceptible to sound change in connected speech. Once a stressed segment is located (based on pitch, intensity and duration of the segment), the lexical match can proceed in a straight forward manner without being concerned about vowel reduction or segment assimilation.

It is clear that a book of phonological rules codifying all the known knowledge in a form suitable for machine analysis of speech would greatly help in the development of speech understanding systems. In the least it could help to pinpoint the relevant rules in English phonology that would significantly help speech analysis. Most rules of this type have been described in articulatory terms, and have neither been translated to acoustical implications nor tested for their general applicability in the English language. Many of the issues raised with regard to the acoustic-phonetic rule "book" also apply to the phonological rule "book."*

A9.6.3 Parametric Variability Resulting From Speaker Characteristics. Most speech recognition systems are trained to work with a single cooperative speaker. Attempts at speaker normalization have not been very successful except for limited cases, such as vowel recognition.

There are at least three types of methods possible for explicity handling the multi-speaker problem in speech recognition:

(1) have separate dictionaries for each speaker;

(2) keep several versions of each entry from different speakers in the same dictionary;

(3) for each speaker, normalize his input speech data according to a transformation whose parameters are determined by sampling his speech (training) and also tune to his speech other parameters used in the segmentation, classification, etc.

The first method requires that the system be trained by each speaker from scratch, with no generalization allowed. Even worse, its storage requirements become prohibitive for a large speaker population.

The second method allows for automatic training on the fly by simply adding to the dictionary when the system makes a mistake (it is assumed in all of these systems that feedback from the user is provided). Its main drawbacks are that the dictionary grows (it may be thought of as the union of the separate dictionaries in method 1) and more drastic, that as the numbers of words and speakers increase, the "words" stored in the dictionary become less separable within their feature space, i.e. more errors will occur. This method can work though for small, well-chosen vocabularies.

Method three gets around the disadvantages of the first two (at the cost of a slightly (?) larger program). It uses just one compact dictionary and allows for training "on the fly." (Alternatively, a new speaker can be asked to utter a few standard sentences from which "his" parameters may be derived.) Its considerable drawback is that no one knows how to do it.

The Vicens-Reddy system was examined with regard to the multi-speaker problem. This system uses method two. In the range for which both multi-speaker and single speaker data is available (54 word vocabularies), recognition of an unknown speaker after training on several (4-9) other speakers as compared to the single speaker case led to a reduction in accuracy from 98% to about 85%, a tripling of computation time and also an approximate tripling of the dictionary size. For larger vocabularies, these degradations can be expected to increase much more.

---

* Often the distinction between the acoustic-phonetic rules and phonological rules is fuzzy. In this report, acoustic-phonetic rules predict the parametric variability in phonetic context and phonological rules predict the parametric variability in morphemic, syntactic and semantic contexts.

The errors were analyzed and blame was placed as follows:*

(a) About 50% were traced directly to initial segmentation "errors" which were so fundamental that the system had no hope of recovering.

(b) About 15% were caused by misclassifying (i.e. mislabeling) one or more segments after the segmentation had been fairly accurate.

(c) About 15% were caused by the differences in the values of the raw speech data, (i.e., the segmentation and classification were more or less accurate.)

(d) The remaining 20% were caused by a compounding of two or more, of a, b, and c.

An attempt was made to improve this system's multi-speaker performance by defining a transformation to be used on the acoustic data corresponding to stressed vowels, an area in which research by Gerstman (1968) has indicated a great deal of consistent inter-speaker difference was to be found. The attempt was to tune the transformation to each speaker by adjusting coefficients. After considerable playing with various types of function, some of which were considerably more complex than the simple linear normalizations which had worked quite well for Gerstman, very little reduction in the inter-speaker variability could be produced.

This failure has two major causes: First, the Vicens data used consists of very wide-band zero-crossing counts which only crudely approximate the high quality formant data of Peterson and Barney which Gerstman used. Also, Peterson and Barney's data was carefully hand-segmented which leads to the second, and more important cause, that the segmentation and representation used by Vicens, although probably the most sophisticated of any speech recognition system yet devised, is still not sufficient to produce better results.

The weakness in Vicens' segmentation is that it is done too locally on the time domain, with almost no context dependency. The basic problem with the representation is that it only allows for a bottom-up analysis: the input acoustic data is transformed through several stages until it is in a particular higher level format; it is then used to search the dictionary, whose elements are also in the same formant. What is needed is a representation and data structure which allows for more feedback down to lower levels to correct "mistakes" and to direct the description and search progress. (This feedback, for a truly powerful speech recognition system, must extend down from syntactic and semantic bases as well.)

These weaknesses are basic to the single speaker problem as well as to that of many speakers. In fact, the errors catalogued above (a, b, c, d) are identical in kind and relative frequency to the errors made when recognizing a single speaker. A single speaker has variability in his acoustic output which can be traced to changes in his emotional condition, his physical health, the time of day, the meaning of his utterance, etc. It is just this variability which makes speech recognition the hard task it is. Variability is increased as we go to several speakers of the same sex, age, locality, and approximate physical characteristics. As each of these restrictions is relaxed, the variability is further increased. It is our belief that differences in the acoustic output caused by the variables occur along the same dimensions; only the magnitude and probability of the differences change.

Thus, the multi-speaker problem is not different than the single speaker one; rather it is the identical problem, only harder. Any technique which improves recognition for a single speaker will lead to better performance for many. Any method for handling several speakers can be used to improve performance for one. (In fact, it is not unreasonable to treat a single speaker as many: when he has a cold, in the early morning, when he is upset, when he is in a hurry, when he is talking about one subject area as opposed to another--in each of these and many other conditions his speech characteristics can change drastically.)

The User Adaptation Problem. Untrained speakers tend to become tense and awkward when they know they are speaking to machines. The resulting effect is that the same sentence will exhibit wide variability at the parametric level in different vocalizations. Thus, in untrained speaker situations the best strategy is to ask the person to relax and speak naturally without making any conscious effort in elocution or enunciation.

In the case of Voice-DM however, we are dealing with a small number of cooperative speakers who, if they need the data badly enough, would be willing to acquire some learned skills and adapt themselves to the situation. In a study to measure the effect of user adaptability on speech recognition, Makhoul (1970) instructed the speakers to change their articulation every time an utterance was incorrectly recognized. Some of the changes in articulation requested from the speaker were: rounding and protruding of the lips and diphthongization, deliberate efforts at voicing and/or frication and proper production of the stop burst.

* Performed by Mr. Lee Erman of Carnegie-Mellon University.

The error rate was reduced from 18.3% to 15% by simple repetition by the speaker after correcting for the articulation. Now the words in the error list were repeated twice (after a brief reminder to the subject of what he supposedly had learned from the first learning session). Those words that were correctly recognized twice in a row were eliminated from the error list. This resulted in a drop in the error rate from 15% to 5.8%. The results indicate that either the errors were random or the speaker was immediately able to change his articulation to effect correct recognition.

The implications for Voice-DM are clear. With some training the speakers can be expected to reduce the error rate 5 to 10% by modifying their enunciation and elocution.

A9.7 The Acoustic Level

Most problems at this level are of an engineering nature, that is, transduction of the changes in air pressure into some digital form either as amplitudes or parameters. The main problems that arise at this level are those of external noise, characteristics of the transducer, and efficient techniques for signal processing.

A9.7.1 The Noise Problem. Input to computers may often have to work in noisy environments, such as computer-room noise, teletype noise, and air-conditioning noise. Very little work has been done to study the effects of noise on machine perception of speech. It would appear that robust techniques that do not degrade the performance of the system significantly in the presence of noise have to be discovered before we can have a reliable speech understanding system.

A9.7.2 The Characteristics of the Transducer. The ready availability of telephone makes it desirable that we attempt to use telephone as the input device to the recognition system. The following types of distortions are known to affect the signal characteristics (Inglis and Tuffnel, 1951; Alexander, Gryb and Nast, 1960; Andrews and Hatch, 1970):

1. Bandwidth limitation. The transmitted band of frequencies is approximately 300-3200 Hz. However, these bounds vary. In addition, a 100 Hz "hole" will sometimes occur somewhere around 2600 cps.

2. Attenuation distortion. The circuit loss over the transmission line results in different levels of attenuation at different frequencies. The loss is relatively flat between 300 Hz to 1100 Hz and rises linearly from 1100 Hz to 3000 Hz. The average difference in loss between 1100 Hz and 2600 Hz is about 8 db.

3. Envelope delay distortion. Phase distortion introduced by the telephone system is measured in terms of the rate of change of phase with respect to frequency, d(0)/dw, which has the dimension of time and is referred to as envelope delay. Frequencies at the low and high ends of the cut-off band exhibit envelope delays of as much a 1 millisecond relative to the delay distortions in the mid-band.

4. Cross modulation (crosstalk). This speech-like noise results from a speech signal occasionally being transferred from another telephone channel.

5. Discretization noise. This is the noise resulting from the digital transmission often used over long distance lines by the telephone system (and not that used before entering the data at a computer). The phone system's digitization uses a 7-bit log PCM encoding.

6. Random noise. Random noise may sometimes be introduced by the digitization process; gaussian noise occurs with all forms of transmission.

There has been no systematic study on the effect of each of these distortions on a speech recognition system. This is due, in part, to the unavailability of working speech recognition systems and, in part, to the apparent inapplicability of results obtained on one recognition system to others. However, it is possible to make some general observations:

1. Effect of bandwidth limitation: Some of the fricatives, e.g., / S, F, TH/ cannot be reliably detected because the primary cues are at frequencies above 3000 Hz.

2. Noise.

   a. Background noise: both random noise and discretization noise are expected to reduce the recognition accuracy a few percentage points as a function of the signal-to-noise ration. In particular, weak sounds such as /F, TH, V, DH, P, H / etc. often cannot be differentiated from the background noise.

   b. Signal distortion: The effect of attenuation distortion and envelope delay are more predictable and could be corrected by the Voice-CS system should it become necessary. What is not known at present is whether a speech recognition system can perform adequately without normalizing for the distortion.

c. Crosstalk: This speech-like noise, is perhaps the hardest of all to correct for. It is analogous to attempting to follow a conversation at a cocktail party; the only way to deal with it is by separating the two conversations. Attempting to correct for crosstalk appears to be beyond the present state-of-art.

3. Characteristics of the Handset: The carbon-button microphone (used in most handsets) is known for its widely varying response characteristics from set to set and from day to day. This makes it difficult for a speech recognition system to normalize for the characteristics of the telephone.

To summarize, telephone input requires the solution of several presently unsolved problems. In particular we need to know what kind of restrictions to the task, language and vocabulary are needed to succeed in building a telephone speech recognition system. To be more specific, can a limited task environment like Voice-CS provide enough restrictions to make it work without any normalization for the telephone distortion? We also need to know what new type of telephone handsets and transmission systems can function conveniently over the present telephone system at a minimal increase in cost to the user. For example can the data rate for digital transmission presently used by the Bell system be increased without excessive cost? What is the cost of sampling twice as often as is presently done? Can the digital data be provided directly to the computer without reconversion to analog form? Answers to these and other such questions can probably be obtained easily but it is not clear whether the answers will be favorable.

A9.7.3 Signal Processing Techniques. Extraction of reliable parameters from the speech signal seems to require sophisticated signal processing techniques that are becoming possible through the use of digital filtering of high-speed FFT hardward devices.

For example, to extract formant parameters, such as those proposed by Schafer and Rabiner (1970), in real time will require computing three 512 point fast Fourier transformations in every 10 milliseconds. There are very few hardward devices that are capable of performing this many operatings and, at present, their cost is prohibitive. More research is indicated in this direction.

With the development of the Fast Fourier Transform, there has been increased interest in digital signal processing techniques. An excellent introduction of this topic can be found in Gold and Rader (1969).

A10. A SIMULATION MODEL FOR PROJECTING THE PERFORMANCE OF SPEECH RECOGNITION SYSTEMS

Given a task, a language for asking questions about the task, we found it desirable to have a system which can predict the performance and indicate the problem areas that are likely to arise in the development of an actual recognition system for the given task. The type of questions that we had hoped to answer were:

1. the effect of the dictionary size and composition on the error rate at the lexical level;

2. the reduction of search that can be attained by careful organization of the lexicon to permit selective search of the lexicon from the knowledge of the phonemes present in the input string;

3. the reduction of search provided by selecting a subset of the lexicon from the knowledge of the semantic situation;

4. the expected number of branches that will survive at each node while performing lexical segmentation of an unending sequence of phonemes without any markings for the boundaries between words;

5. the reduction of combinatorial exploration in obtaining lexical match of segments with multiple labels;

6. the effect of increase in vocabulary on the observed phonemic ambiguity among the words of the vocabulary.

It was observed by Newell and Reddy that the phonemic ambiguity analysis system that is available at Carnegie-Mellon University could be used with appropriate modifications to provide answers to most of these questions. Here we will present the details of this phonemic ambiguity analysis system and how it was used to answer the questions raised above.*

A10.1 The Model

From available linguistic knowledge we can construct a representation for a word in terms of a string of phonemes, and a representation for each phoneme in terms of a set of articulatory features. These articulatory features are related to the parametric representation. Figure A10.1 provides a crude estimation of the importance of each of these articulatory features. The value of -1 indicates that this articulatory feature is not relevant and should be ignored.

From this representation of phonemes, one can calculate the similarity between any two phonemes by calculating the differences of the features between the two phonemes, multiplying them by the relative weights, and suitably normalizing them. Figure A10.2 gives one such table. Accepting this table for the moment, we can calculate the similarity between any two words, i.e., any two strings of phonemes. This is done by multiplying together the similarities of corresponding phonemes. For instance, if the two words were "DECtape" and "magtape" we would calculate:

```
DECTAPE =  D   E   K   T   E   P
           |   |   |   |   |   |
MAGTAPE =  M  AE   G   T   E   P
```

$$.71 * .79 * .9 * 1 * 1 * 1 = .50$$

The use of a scale of 1 for identical and 0 for completely dissimilar, along with a multiplicative combination of scores, reflects the decision criteria in a match that all components must be the same.

The example above had components in one to one correspondence. A major complication is that phonemes influence eachother in speech, so that two highly similar adjacent phonemes are sounded as one. Thus, phoneme strings do not have to match in length to represent the same word. A calculation rule can be formed which reflects this consideration in a gross way and allows similarity scores to be formed between words of different phonemic length. The phonemes that exist in both words are put in correspondence, and then account taken that the extra phonemes could have been assimilated. The calculations below for "sit" versus "slit" and "sit" versus "split" show the scheme.

```
SIT   S          I   T
      |          |   |
SLIT  S — L  —   I   T
```

$$1 * \max(P_{SL}, P_{LI}) * 1 * 1$$

$$= \max(.51, .79) = .79$$

```
SIT   S                  I T
      |                  | |
SPLIT S — P  —  L  —     I T
```

$$1 * \max(P_{SP}, P_{PL}) * \max(P_{PL}, P_{LI}) * 1 * 1$$

$$= .66 * .79 = .52$$

For any given pair of phoneme strings, the model uses the highest similarity score over all possible associations of phoneme pairs.

* We would like to acknowledge the special contributions of L. Erman, G. Goodman, D. McCracken and R. Neely in formulating and obtaining the results of this model.

```
;THE QUALITIES ARE: VOCALITY; NASALITY; CONSONT LOC; AFT -> FORE;
;    DOWN -> UP; FRICATION; LIQUID; BURST; AND OPENNESS.
;VOCALITY: 0 FOR VOICELESS, 1 FOR VOICED
;NASALITY: 0 FOR NOT NASAL, 1 FOR NASAL
;FRICATION: 0 FOR NO FRICATIVE, 1 FOR F-TYPE,
;    9 FOR SH-TYPE, 10 FOR S-TYPE
;LIQUID: 0 FOR NOT LIQUID, 1 FOR A LIQUID
;BURST: 0 FOR NO BURST, 1 FOR A BURST
;OPENNESS: 0 TO 10 FOR CLOSED TO OPEN

5
90
90
90
```

;	VOICE	NASAL	CNS LC	B-F	H-L	FRIC	LIQ	BRST	OPEN	DUR
10	5	1	1	2	2	1	4	1	8	2
--	0	0	-1	-1	-1	0	0	3	0	5
P	0	0	10	-1	-1	0	0	4	0	5
B	1	5	10	-1	-1	0	0	4	0	5
T	0	0	6	-1	-1	0	0	4	0	5
D	1	5	6	-1	-1	0	0	4	0	5
CH	0	0	5	-1	-1	9	0	4	0	5
J	1	3	5	-1	-1	9	0	4	0	5
K	0	0	2	-1	-1	0	0	4	0	5
G	1	5	2	-1	-1	0	0	4	0	5
F	0	0	9	-1	-1	2	0	0	0	5
V	1	5	9	-1	-1	2	0	0	0	5
TH	0	0	8	-1	-1	2	0	0	0	5
DH	1	4	8	-1	-1	2	0	0	0	5
S	0	0	6	-1	-1	10	0	0	1	9
Z	1	0	6	-1	-1	10	0	0	1	9
SH	0	0	4	-1	-1	9	0	0	1	9
ZH	1	0	4	-1	-1	9	0	0	1	9
H	0	0	0	-1	-1	6	0	0	1	7
M	1	10	10	-1	-1	0	0	0	2	5
N	1	10	6	-1	-1	0	0	0	2	5
NG	1	10	2	-1	-1	0	0	0	2	5
W	1	7	10	-1	-1	0	1	0	2	2
R	1	0	7	-1	-1	0	1	0	3	2
L	1	0	6	-1	-1	0	1	0	3	2
Y	1	0	4	-1	-1	0	1	0	2	2
OO	1	7	1	1	8	0	0	0	4	9
U	1	5	2	2	6	0	0	0	3	5
O	1	0	1	1	6	0	0	0	6	8
AW	1	0	0	0	1	0	0	0	8	9
AA	1	0	0	0	0	0	0	0	8	9
A	1	0	5	5	6	0	0	0	6	5
AR	1	0	6	6	8	0	0	0	6	7
AE	1	0	7	7	3	0	0	0	6	7
E	1	0	8	8	6	0	0	0	5	9
I	1	0	9	9	8	0	0	0	4	5
EE	1	0	10	10	10	0	0	0	4	9
SW	1	0	5	-1	-1	0	0	0	1	3

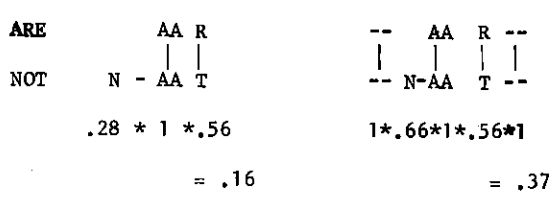
Figure A10.1: Phoneme Feature Weights

```
      --
      |  P
--   100 |  B
P     99 100 |  T
B     89 90 100 |  D
T     99 96 86 100 |  CH
D     89 86 96 90 100 |  J
CH    90 86 76 90 80 100 |  K
J     82 78 84 82 88 92 100 |  G
K     99 92 82 96 86 88 80 100 |  F
G     89 82 92 86 96 78 86 90 100 |  V
F     95 93 83 91 81 85 77 87 77 100 |  TH
V     85 83 93 81 91 75 83 77 87 90 100 |  DH
TH    95 92 82 92 82 86 78 88 78 99 89 100 |  S
DH    86 83 91 83 91 77 85 79 87 90 98 91 100 |  Z
S     71 66 56 70 60 78 70 66 56 73 63 74 65 100 |  SH
Z     66 61 61 65 65 73 75 61 61 68 68 69 70 95 100 |  ZH
SH    72 65 55 69 59 79 71 69 59 72 62 73 64 97 92 100 |  H
ZH    67 60 60 64 64 74 76 64 64 67 67 68 69 92 97 95 100 |  M
H     79 68 58 72 62 76 68 76 66 75 65 76 67 86 81 89 84 100 |  N
M     66 65 75 61 71 51 59 57 67 66 76 65 74 55 60 54 59 57 100 |  NG
N     66 61 71 65 75 55 63 61 71 64 74 65 74 59 64 58 63 61 96 100 |  W
NG    66 57 67 61 71 53 61 65 75 60 70 61 70 55 60 58 63 65 92 96 100 |  R
W     59 58 68 54 64 44 52 50 60 59 69 58 67 48 53 47 52 50 87 83 79 100 |  L
R     58 54 54 56 56 46 48 52 52 57 57 58 59 50 55 49 54 52 69 71 67 82 100 |  Y
L     58 53 53 57 57 47 49 53 53 56 56 57 58 51 56 50 55 53 68 72 68 81 99 100 |  OO
Y     66 59 59 63 63 55 57 63 63 62 62 63 64 57 62 60 65 63 74 78 78 87 89 90 100 |  U
OO    45 35 45 39 49 31 39 43 53 38 48 39 48 49 54 52 57 53 64 68 72 57 61 62 56 100 |  O
U     63 54 64 58 68 50 58 62 72 57 67 58 67 52 57 55 60 62 79 83 87 72 80 81 75 75 100 |  AW
O     38 28 28 32 32 24 26 36 36 31 31 32 33 38 43 41 46 46 43 47 51 36 54 55 49 71 62 100 |  AA
AW    20  9  9 13 13  5  7 17 17 12 12 13 14 23 28 26 31 29 24 28 32 17 35 36 30 44 31 69 100 |  A
AA    20  9  9 13 13  5  7 17 17 12 12 13 14 23 28 26 31 29 24 28 32 17 35 36 30 42 29 67 98 100 |  AR
A     44 38 38 42 42 34 36 40 40 41 41 42 43 36 41 37 42 40 53 57 55 46 64 65 57 53 62 82 51 49 100 |  AE
AR    40 35 35 39 39 29 31 35 35 38 38 39 40 41 46 40 45 43 50 54 50 43 61 62 52 58 51 79 48 46 89 100 |  E
AE    40 36 36 38 38 28 30 34 34 39 39 40 41 40 45 39 44 42 51 53 49 44 62 61 51 45 46 74 55 53 84 87 100 |  I
E     44 41 41 41 41 31 33 37 37 44 44 45 46 51 56 50 55 45 56 56 52 49 65 64 54 60 53 69 42 40 75 78 79 100 |  EE
I     60 58 58 56 56 46 48 52 52 61 61 60 61 50 55 49 54 52 73 71 67 66 80 79 69 61 62 50 19 17 68 71 64 77 100 |  SW
EE    52 51 51 47 47 37 39 43 43 52 52 51 52 57 62 56 61 51 66 62 58 59 71 70 60 62 47 47 20 18 53 64 57 78 85 100 |
SW    80 74 74 78 78 70 72 76 76 77 77 78 79 72 77 73 78 76 73 77 75 74 76 77 85 53 72 46 27 27 56 51 50 53 68 59 100
```

A note of caution. These values were generated by a program which used a weighted similarity measure based on the distinctive features of phonemes. The weights were chosen manually based on empirical observation and may be incorrect in some cases.

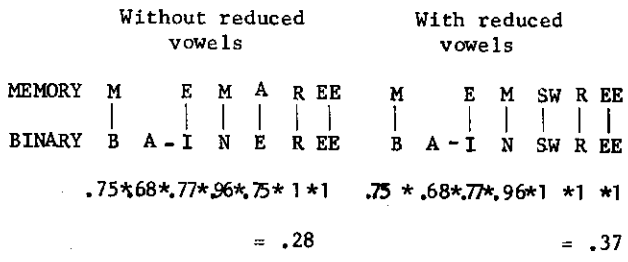
Figure A10.2: Phoneme Similarity Matrix
(Numbers represent percentage similarity)

It is important that the phonemes that bound the word be taken into account, since assimilation can occur with these as well. Taking words in isolation (i.e., still putting off questions of continuous speech), we can bound each word by a silence phoneme (--). Thus, the similarity between "are" and "not" goes as follows:



Bounding by the silence phoneme increases the similarity between "are" and "not" from .16 to .37, thereby more accurately predicting the probability of confusion between these two words.

Another important effect is the reduction of vowel duration and intensity when it occurs in a non-stressed position. Again this can be taken into account in the matching algorithm in a rough way. The following computation illustrates how the similarity score between "memory" and "binary" increases from .28 to .37 when the vowel reduction is taken into account by replacing the unstressed vowels by the neutral vowel schwa (/SW/).



There are other effects in the actual programs that the above calculations do not take into account. Of course, there are several aspects of phonetic realization, such as the changes in duration with phonemic context, that neither the calculations above nor the actual programs take into account. But this will do for a rough picture.

A10.2 Validation of the Model

Acutally, of course, these calculations, both for the phonemic similarity and the derived word similarities, constitute a crude theory of the recognition process. We should be able to evaluate how well it predicts existing performance of recognition systems. Unfortunately, there is little data of the requisite quantity and quality with which to make the comparison. However, we do have adequate data from the Vicens-Reddy system on a vocabulary of 54 words spoken in isolation. This same vocabulary was also used in two other investigations (Gold; Bobrow and Klatt), though the data is not published.

For the Vicens program we have several runs on the 54 words, in which each word was said and the program made a recognition against the full dictionary of 54 words. The runs are described in Figure 10.3. We know exactly which errors were made for what words. Thus, we can compare the model's similarity scores for error pairs with the similarity score for non-errors. We certainly should not expect the errors to be all those and only those which are above a given threshold, since the process is inherently statistical. But we should expect the errors to favor high scores strongly.

Two comparisons are worth presenting. In the first (Figure 10.4) we see the distribution of rank orders of the errors. For instance, there was an error of "five" for "divide," but there was another word, "byte," whose similarity score with "divide" was higher than that of "five" (50 compared to 49). Thus, a tally was made for an error at rank 2. The advantage of the rank order is that it compares the competitors for a given word, independent of deficiencies of the similarity model in comparing between quite different situations. As one can see from Figure 10.4, the errors cluster toward the high ranks,* though with a scattering of ranks all the way down to the last ranks. (There are exactly 53 occurrences of each rank.)

The second display of the errors (Figure 10.5) gives the distribution of absolute scores for the errors against the distribution of absolute scores for the entire 54x53 matrix of comparisons. The figures are percentages in each 10 point category, taking into account the multiplicity of occurrences of errors (e.g., "core" was given for "four" 9 times). The ratio of these two frequencies, suitably normalized to account for the total number of runs in the sample, gives an estimate of the probability of error given the similarity score. These values are also shown in Figure 10.5. This ratio should be relatively independent of the particular vocabulary used, though of course not of the details of the Vicens-Reddy program. Again, we see the same effect as in Figure 10.4: that there is a relatively high probability even if the similarity is low. In both Figure 10.4 and 10.5 we have segregated the three types of runs

* The small secondary peak in the range 16-25 for class C in Figure 10.4 appears to indicate that some of the important mechanisms are yet to be captured by the model.

Data Type	Recording Quality	Number of Speakers	System Adaptation	Number of Runs	Number of Trials	Number of Errors	Error Rate
A	Good	1	Poor to Good	8	432	14	3.2%
B	Medium (15 db S/N)	10	*Good	5	270	33	12.2%
C	Medium (15 db S/N)	10	**Poor	13	702	189	26.9%

* Class B consists of runs made after first training the system on other speakers.

** Class C consists of runs made while first training the system.

Figure A10.3: The three types of data from the Vicens program used for calibration of the Model. (See pages 129, 135 and 136 of Vicens (1969) for further details of this data).

		Class A Data	Class B Data	Class C Data	Total Data
		% of 14 Total Errors	% of 33 Total Errors	% of 189 Total Errors	% of 236 Total Errors
Rank in Similarity Ordering	1-5	78.6	72.7	53.4	57.6
	6-10	7.1.	9.1	11.1	10.6
	11-15	7.1	9.1	7.4	7.6
	16-20	0	9.1	11.1	10.2
	21-25	0	0	10.1	8.1
	26-30	0	0	2.1	1.7
	31-35	0	0	2.1	1.7
	36-40	0	0	1.1	0.8
	41-45	0	0	0.6	0.4
	46-50	7.1	0	0.6	0.8
	51-54	0	0	0.6	0.4

Figure A10.4: Distributions of rank orders of errors.

Distribution of Absolute Scores

Decade Range	Entire 53x54 Matrix	Class A Errors	Class B Errors	Class C Errors	Total Errors
90-99	0.0%	0.0%	0.0%	0.0%	0.0%
80-89	0.1	21.4	9.1	6.9	8.1
70-79	0.3	21.4	6.1	2.6	4.2
60-69	0.4	0.0	3.0	3.2	3.0
50-59	1.5	0.0	12.1	13.8	12.7
40-49	3.8	28.6	24.2	15.3	17.4
30-39	7.5	7.1	21.2	13.2	14.0
20-29	18.9	14.3	21.2	32.8	30.1
10-19	29.8	0.0	3.0	5.3	4.7
0-9	37.7	7.1	0.0	6.9	5.9

Estimate of Probability of Error Given Similarity Score

Decade Range	Using Class A Errors	Using Class B Errors	Using Class C Errors	Using Combined Errors
90-99	1.0*	1.0*	1.0*	1.0*
80-89	.375	.60	1.0	.73
70-79	.094	.10	.096	.096
60-69	0	.033	.077	.045
50-59	0	.038	.095	.055
40-49	.009	.030	.041	.029
30-39	.001	.013	.018	.012
20-29	.0009	.005	.018	.010
10-19	0	.0004	.002	.001
0-9	.0002	0	.002	.001

* No pairs in the 54 word lists used to generate these data fell into the 90-99 decade; these probabilities are therefore arbitrarily set to 1.0.

Figure A10.5: Distribution of absolute scores for errors against entire 53x54 matrix and resulting estimate of probability of error given similarity score.

as described in Figure 10.3. We note that the better the data (high quality and high performance) the better the model seems to fit the error data.

The above attempt at validation is not only rough, but limited to a particular program. Changes in the logic, etc., will make differences in the performance. However, the behavior of each of three different programs (Vicens-Reddy, Gold, Bobrow and Klatt) is somewhat comparable thus indicating that the gross performance figures can be taken as indicative of the state of the art. Thus, though we must interpret the results with care, it appears useful to use the model to explore various aspects of the Voice-CS program.

A10.3 Conclusion

The model described here has been used to answer the questions raised at the beginning of this appendix. These analyses and the obtained results are presented in detail in Appendix 9. For the most part, they agree with one's intuitive notions.

A11. PHONEMIC ANALYSIS OF A FREE ENGLISH SENTENCE*

This appendix contains an analysis of a telephone recording of:

> Ha, ha.. Ok.. um.... let's see.. wh. What things would I want to know right off hand--I'd like to know how many users are on the line? On.. are on the system that's obvious.

This fragment was taken from an actual protocol for Voice-CS (Section A6). The analysis is a hand simulation of the phoneme recognizer of Forgie and Forgie (1959). Some assumptions are made with regard to extending that scheme to handle telephone bandwith and continuous speech.

Column 1 of the analysis is a guess at what sounds are really present on the recording (section A11.1 contains the codes used in the first three columns). Next is the segmentation which the simulated program would make. The third column shows the hypothetical results of a phoneme recognizer; it contains an ordered list of phonemes with a confidence score for each (on an arbitrary 0-100 scale). The list is arbitrarily truncated to eliminate less likely candidates. The last column is the English transcription, approximately lined up with the segmentation.

It is expected that the machine would actually make more errors than indicated. Almost all errors are related to segmentation problems, or, conversely, if the segmentation is correct, then the phoneme classification is very likely to be correct also.

A11.1 Glossary of Words Used in the Analysis

Segment Codes

S	Silence
**	Noise (non-speech)
FR	Fricative
VF	Voiced Fricative
H	Aspiration
PL	Plosive Burst (Voiceless)
VP	Voiced Plosive Burst
V	Vowel
SV	Semi-Vowel
N	Nasal
VS	Voiced Silence

Phoneme Codes

Vowels		Consonants	
AE	bad	B	
AH	father	CH	church
AI	bite	D	
AW	awe	DG	jump
AY	bay	DH	then
E	bet	F	
EE	beet	G	
ER	bird	H	
I	bit	K	
L*	able	L	
O	open	M	
OO	boot	N	
OU	bout	NG	
U	put	P	
UH	but	R	
U*	about	S	
		SH	
		T	
		TH	with
		V	
		W	
		X	_on
		Y	yet
		Z	
		ZH	vision

* Produced by J. W. Forgie and C. D. Forgie

Transcription	Segment Code	Phoneme Scores	English
laugh	**		Ha. Ha...
	V	← We have omitted classification of these sounds	
	S	since we have no experience with them, but a	
	**	real recognizer would have to deal with	
	h	them at this level.	
	V		
	V		
	S		
	**		
	V		
	S		
O	V	O 83, AH 77, ER 64, AE 67, OU 56, UH 51, OO 43	OK..
	S		
K	PL	K 88, T 74, P 53, G 40, D 39, B 32, S 20, SH 26, H 22	
AY	V	AY 71, EE 65, I 63, E 58, AI 45, AH 42, ER 40	
	S		
UH	V	UH 81, AE 75, ER 72, AH 66, OU 54	um.
M	N	M 85, N 67, NG 43	
Noise	**		
	S		
	**		
	S		
L	N	L 55, W 50, R 48, M 46, N 45	let's
E	V	E 75, I 68, AE 52, U 56, UH 65, ER 54	
T	S	T 70, K 65, P 58 } Alternate analysis CH 68	
S	FR	SH 63, H 56, S 42, F 41, TH 41	
EE	V	EE 72, AY 70, AI 65, E 65, I 62, AE 58	see..
W	SV	W 60, L 32, N 24	Wh.
L*	V	L* 70, AH 62, UH 60, FR 56, E 47, O 44, AI 42, OU 42	
	S		
W	N	W 55, M 48, L 38, N 35	What
UH	V	AH 58, U 56, UH 54, AE 50, I 48, E 44, AI 42	
	S		
T	PL	T 68, K 62, D 52, G 47, P 43, B 36, TH 34, F 30	things
TH	-		
I	V	I 76, AY 74, E 72, EE 72, ER 56, AI 50	
NG	-		
Z	FR	S 88, SH 63, Z 56, ZH 42, F 38, TH 38	would
U	V	U 70, E 65, UH 58, I 56	
D	-		
AI	V	U 65, I 60, AI 54, ER 52, E 48	I
W	SV	W 50, M 43, L 30	
UH	V	UH 68, E 65, U 55, AE 52	want
N	N	N 58, NG 45, M 40, L 30	
T	-		
U*	V	UH 65, U 58, E 49	to
N	N	N 58, NG 50, M 42, L 25	
O		} These three phonemes would not be segmented by our	know
R	V	present techniques. We would probably call the whole	
AI		thing AI.	right
	VS		
D	VP	D 72, G 65, B 46, T 48, K 48, P 37	
AW	V	AW 68, OU 65, UH 62, AH 60, U 56	off
F	S		
H	h	H 80, T 50, K 48, P 36	hand
AE	V	AE 66, E 66, UH 62, AH 60, AW 60, U 59, OU 52	
N	N	N 50, M 42, NG 40	
D	S		
Click	FR	SH 80, S 70, H 50, ZH 42, Z 40, F 36, TH 30	
	PL	T 70, K 65, P 50, D 40, G 36, B 32	
	S		

Transcription	Segment Code	Phoneme Scores	English
AI	V	UH 60, E 56, U 56, AH 52	I'd
D	VS		
L	N	L 70, M 50, R 50, N 40, W 40	like
AI	V	AI 88, AE 70, E 74, UH 67, AH 64	
	S		
K	PL	K 50, T 47, P 42, D 26, G 20, B 21	
	S		
T	PL	T 70, K 64, P 40	
U*	h	This appears to be an unvoiced vowel for which we have no analysis.	to
	S		
N	N	N 70, NG 45, L 40, M 36	know
O	V	OU 68, O 64, AW 56, U 56	
H	VF	DH 50, V 40, H 40, ZH 34, Z 30	how
OU	V	OU 70, O 60, AW 58, U 52	
M	N	M 60, N 56, NG 45	many
E	V	AI 76, AE 65, E 62, UH 55	
N	N	N 70, NG 60, M 48	
EE	V	AI 72, EE 69, AY 65, I 60	
Y	SU	Y 70, L 50, W 30 Good chance of missing this segment	users
OO	V	U 60, UH 58, I 50	
Z	VF	Z 80, S 70, ZH 68, SH 60, DH 40, V 40	
ER	V	E 70, I 67, UH 55, ER 52	
Z	FR	S 80, Z 70, SH 68, ZH 60, TH 40, F 40	
ER	V	ER 74, AW 74, AE 71, AH 62	are on the
AH	V	AH 88, UH 72, ER 66, AW 50	
N	N	N 70, M 59, NG 54	
DH	VP	D 56, G 52, DH 48, V 48, B 44	
U*	V	O 68, OU 65, UH 60, U 55	line..
L	SU	L 50, W 45, R 40	
AI	V	AI 85, AH 70, AY 60	
N	N	NG 42, N 36, M 40	
	S		
breath?	h		
	S		
X	VP	X 70, G 60, B 50, D 40	on..
AH	V	AH 90, AE 78, UH 68, E 60, O 68	
N	N	N 85, M 45, NG 42	
	S		
X	PL	K 50, H 50, P 45, T 35	are on the
ER		} We would not segment this	
AH	V	AH 70, AW 60, UH 50 string with present techniques.	
N			
TH	S		
S	FR	H 82, SH 75, S 58, K 47, T 45	system.
I	V?	Another unvoiced vowel.	
S	FR	H 82, SH 75, S 58, K 47, T 45	
	S		
T	PL	T 88, K 80, P 55, D 50, G 48, B 35	
U*	V	UH 82, AE 70, E 68, U 64, AW 64, O 58	
M	N	M 80, N 65, NG 50, L 40	
	S		
	**		
	S		
DH	VP	D 82, G 76, DH 72, V 68, B 64	that's
AE	V	UH 82, AE 77, E 74, U 74	
T	S	T 70, K 65, P 45 } Alternative Analysis CH 75, DG 60	
S	FR	SH 82, S 73, TH 50, F 50	
AH	V	AH 80, O 76, ER 70, AE 62, UH 60, OU 58	obvious.
V	VS		
Y	V	EE 76, I 66, UH 60, E 50, AH 48	
U*			
S	FR	S 76, SH 60, F 50, TH 50, Z 42, ZH 40	
	S		

A12: ALTERNATIVE MANAGEMENT SCHEMES*

There are about 12 factors or parameters that appear to be especially significant in determining the possible structures of projects. These factors or parameters are:

1. The projected duration of the project.
2. The number of phases into which the project is divided.
3. The number of contractors participating in the project.
4. The number of functions or tasks the speech system is designed to handle.
5. The degree of simplicity or of complexity and sophistication of the system.
6. The amount of syntactic and/or semantic support derived from the function, task, or situation.
7. The amenability of the function or task for tuning the system for individual speakers and/or for training the speakers.
8. The extent to which research is involved in the program (versus the extent to which the program is essentially a development program).
9. Whether the project is managed with conventional techniques or with a high degree of involvement of the ARPA Network.
10. The disposition of administrative and technical control of the project.
11. The extent to which specialized supporting activities and/or specialized technical arrangements, such as the mounting of an organized data-collection program and the joint use of specialized measurement or analysis equipment, is/or involved.
12. Whether or not substantial involvement of a significant number of long-experienced speech researchers is made a fundamental tenet of management of the project.

Given so many factors and the possibility of having several degrees or treatments of each, one might well suppose that the only possibility of thinking about the matter is to find or prepare a multidimensional artificial intelligence program that will find eigenvalues in unquantifiable situations. However, the problem may not be as bad as all that. If we reserve the eleventh and twelfth factors for subsequent discussion, we may be able to reduce the whole business to the description of about a dozen project structures that seem plausible or interesting. In the following paragraph, I shall set forth descriptions of ten project structures. Doubtless you will find others that you consider more plausible or more interesting If you do, this effort will have been successful.

Ten Possible Project Structures

Figure 12.1 shows ten possible project structures, A through H, each one being defined with reference to the first ten of the twelve factors listed earlier. In the table, I have left room for two more structures to be invented by the reader. In the following paragraphs I shall try to explain what the table means to me.

First, project structures A through G were the only ones that suggested themselves to me as being plausible or interesting, but I was missing an obviously good bet in H, which was suggested by Allen Newell during a telephone conversation. I may not define H very well, but I think I can explain its significant feature, and I shall come to that in due course.

Structure A is a three-year, two-phase project with a single contractor. It would be charged with the responsibility of building a system to handle one or two functions. ("Functions" are approximately the same as the "tasks" in terms of which we thought would be to operate a highly constrained data base. Another would be to operate a system for receiving and organizing "debriefings" from pilots returning from missions.) The system and its functions would be simple. The functions and the situation would provide a lot of semantic and syntactic support. It would be possible to tune the system for each individual speaker and to train the speakers. (I imagine that a system would be "tunable" if there were no more than, say, 100 speakers who used it.) The project would be essentially a development project, though this does not rule out all research. Being a single-contractor project, it would be operated in a conventional way. Administrative and technical control would be vested in the contractor -- except, of course, for the basic fiscal and veto control that ARPA would of course retain. This project does not seem interesting to me, but it does seem plausible. It would probably turn out to be

* Written by J. C. R. Licklider.

	A	B	C	D	E	F	G	H
Years	3	3	3	5	5	5	5	5
Phases	2	2	2	3	3	3	3	3
Contractors	1	3	5-7	1	3-4	3-7	6-8	6-8
Functions	1-2	3	2-3	2	2-3	2-5	3-6	1-6
Simple/Complex	S	S	S	M	MC	S-MC	S-MC	S-MC
syntactic-semantic support	Hi	Hi	Hi	MHi	M-MHi	M-MHi	M-MHi	M-MHi
tunability/trainability	Hi	Lo-Hi	Lo-Hi	Lo-Hi	Lo-Hi	Lo-Hi	Lo-Hi	Lo/Hi or Lo-Hi
R and D/development	D	D	D	D+R	D+R	D+R	D+R	D+R
conventional/network	C	C_{+n}	C_{+n}	C	C+N	C+N	C+N	C+N
administrative and technical control	C	3-way C	ARPA or prime	C	circle	ARPA or prime	ARPA + circle	ARPA + coord

Figure 12.1: Alternative Project Structures
(See text for explanation.)

successful in a minor way, but I doubt that it would demonstrate a sufficiently impressive performance in speech recognition and "understanding" to make ARPA feel that it had, indeed, done a major or significant thing. But of course I could be wrong about this.

Structure B is for another three-year, two-phase project. There would be three contractors, coordinate with one another. Each would build a system to fulfill a single function or task. All the functions would be simple. All would have high syntactic and semantic support. At least one of the systems would be for a situation in which a high degree of tunability and trainability would be possible, and at least one of them would be for a situation of just the other kind. It would be interesting to compare the two classes, insofar as one can do that with only one exemplar of one and two exemplars of the other, in terms of feasibility as reflected through progress and success. This project, as indeed all the three-year projects, would be essentially a development project. It would be managed basically through conventional techniques, but such use as turned out to be convenient would be that of the ARPA network. Each contractor would do his own thing in his own way, and a certain amount of competition would prevail. However, ARPA's contractors are by nature friendly, and it would be expected that there would be some cooperation and some interchange also. This project seems a bit more interesting to me than the first one, but this one too suffers from the fact, which I think is probably overriding, that it will be very difficult to do enough in three years to convince the world that a truly significant achievement has been made.

Structure C is a three-year, two-phase project with five to seven contractors. Among them, they try to develop systems covering two or three functions. All are simple with high syntactic and semantic support. As was the case in structure C, however, at least one has high tunability-trainability and at least one low. All are essentially development projects. All are handled more or less conventionally, with some help from the network. Administrative and technical control is centralized, either in ARPA or in a prime contractor. The effort to achieve cooperation and fitting together of advances, products, sub-systems, and the like is major. This structure may seem a bit interesting to me, but I think it is too complicated for a three-year project, and I have to say that, of the three-year structures, I like B the best.

Structure D is the first of the five-year three-phase projects. There is a single contractor. He undertakes to develop systems to handle two somewhat dissimiliar functions. Both are moderately complex, and both have moderately high -- but not very high -- syntactic and semantic support. One has low tunability and trainability, the other high. The central effort is development, but it is supported by a significant amount of research. With only a single contractor, the management methods are essentially conventional. Administrative and technical control are vested in the contractor. Given a good contractor, this might be a very good project structure. The main trouble would be getting enough real competence in speech and language to come together within the limits of a single contracting organization. This whole thing is going to depend, it seems to me, upon getting truly high competence, and a fair amount of it, to bear upon the problem.

Structure E differs from structure D in involving three or four contractors, rather than one, which would make it possible to undertake to handle more different functions -- but that possibility is not exploited. The functions and the systems are perhaps somewhat more complex in E. The syntactic and semantic support is moderate to medium high. Again, both ends of the tunability and trainability scales are represented. Again there is a significant amount of research to support the development. This time the network comes in for a significant role. The administrative and technical pattern is essentially that of a circle of contractors, all coordinate. ARPA exercises fiscal and, of course, technical veto power, but the essential pattern is that of a somewhat competitive, somewhat cooperative association, a loose federation, of contractors -- something like a smaller version of the overall ARPA contractor community in computer science and engineering. (Or perhaps ARPA does not think of it in quite those terms.) I think that the project structure is very interesting and that it would be a lot of fun to participate in it. I doubt, however, that it would be as productive as the next one, F.

Structure F would embrace the same number of contractors or perhaps more. It might cover a wider gamut of functions and a wider range on the simplicity-complexity scale. It would be about the same in terms of the next variables, on down to administrative and technical control. Here, however, it is assumed that ARPA or a prime contractor would exercise diligent and rigorous administrative and technical control and make every effort to weld the several contractors into an efficient problem-solving and development organization. Significant use would be made of the ARPA Network in this effort. I like this project structure very much. I think it would have a good chance of succeeding in a significant way, and I think

that -- in the process of trying to succeed in a significant way -- it might learn some very interesting techniques about management with the aid of a computer network.

Structure G differs from E and F in admitting more contractors and more functions. To a considerable extent, it operates in the manner of the circle described in connection with E, but, being larger and therefore requiring more administrative and technical control, it relies upon ARPA -- either upon a person of Larry Robert's caliber in ARPA or upon such a person brought into the situation through contract or consultantship but reporting rather directly to ARPA -- it relies upon such a person for a greater degree of administrative control and technical direction than a loose confederation would have. This is an effort to achieve a fair degree of direction without having a prime contractor. In this plan, all the contractors would be coordinate. That seems to me to be in some ways a good feature, for I am not sure how effectively ARPA contractors would work if one of them were the prime contractor and the other sub-contractors. On the other hand, it would be asking quite a bit of one man, especially if he were not actually a member of ARPA, to exercise the required amount of control over a circle of contractors.

Structure H is based on a suggestion by Allen Newell that I consider to be an excellent one. It is that the "circle" kind of organization be supplemented by something that might be called a "cooperation expediter." Assumed that there is a circle of contractors and, in addition, a single person -- not so much a director as a facilitator or expediter of cooperation -- who has control of enough money to put into effect on short notice various cooperative or supporting plans that are agreed upon by him and two or more members of the circle. For example, he might convene a meeting of experts on some phase of speech analysis. He might let a small contract for the quick implementation of a special measuring device. In any event, this structure might involve a considerable number of contractors. It might work with almost any reasonable number of functions. It would probably make a great amount of use of the network. I like it very much, but on reflection I find myself coming to the conclusion that it is very important for this enterprise to be successful and that the best chance of making it so is to impose, in the project structure, a considerable amount of administrative and technical control. I think that the psychological effect of such control would be good rather than bad in the development parts of the enterprise. I think that the more or less significant incompatibility of administrative and technical control, on the one hand, and creative research, on the other, might be avoided by deliberately exempting the research parts of the program from hard-driving management.

In Figure 12.2, I have shown schematically what I have in mind with respect to two-phase and three-phase projects. In the three-year projects, I suppose that there would be time for two overlapping two-year development efforts. There would be room for a small amount of research. It is shown feeding into the development project, and the first development project is shown feeding into the second. The square with vertical lines above the development bars are to represent periods during which the prototype systems are exercised. I think it is very important to go through the development and exercising functions once before the final time, and there is certainly not room in a three-year schedule for more than two such phases. Indeed, I think they would have to overlap more or less as I have shown.

In a five-year project, there is probably room for three development phases, but, again, they must overlap. In the five-year schedule, there is of course more room for research and more chance for it to feed into the development efforts. Accordingly, I would put the ratio of research to development at about 10% in a three-year project and at about 30% in a five-year project.

Figure 12.3 shows in a very schematic and probably not very helpful way four different kinds of project "management." The trouble with the single-contractor arrangement, as I have suggested, is that it will be very difficult to get enough competents together under one roof. The trouble with the circle of contractors is, of course, that such an arrangement is delightful for research but probably too relaxed and disorganized for development. The prime-contractor/subcontractor arrangement is standard in industry for development programs that require an array of competences that a single contractor cannot provide, and I think that the only problem with it in the present context is that it will be necessary to use research people in a program that leans heavily toward development if ARPA is to bring off a truly significant accomplishment in the short time projected -- in either of the short times projected. Actually, this project would be quite different from many development projects in that it would involve, essentially, a synthesis of research ideas and findings from the frontier of knowledge. Such a synthesis is hard to distinguish from research. It would certainly be interesting. It seems that it should be possible to overcome the researcher's traditional distaste for "businesslike" management methods in such a situation. Therefore I like the prime-contractor/subcontractor schema. I like,

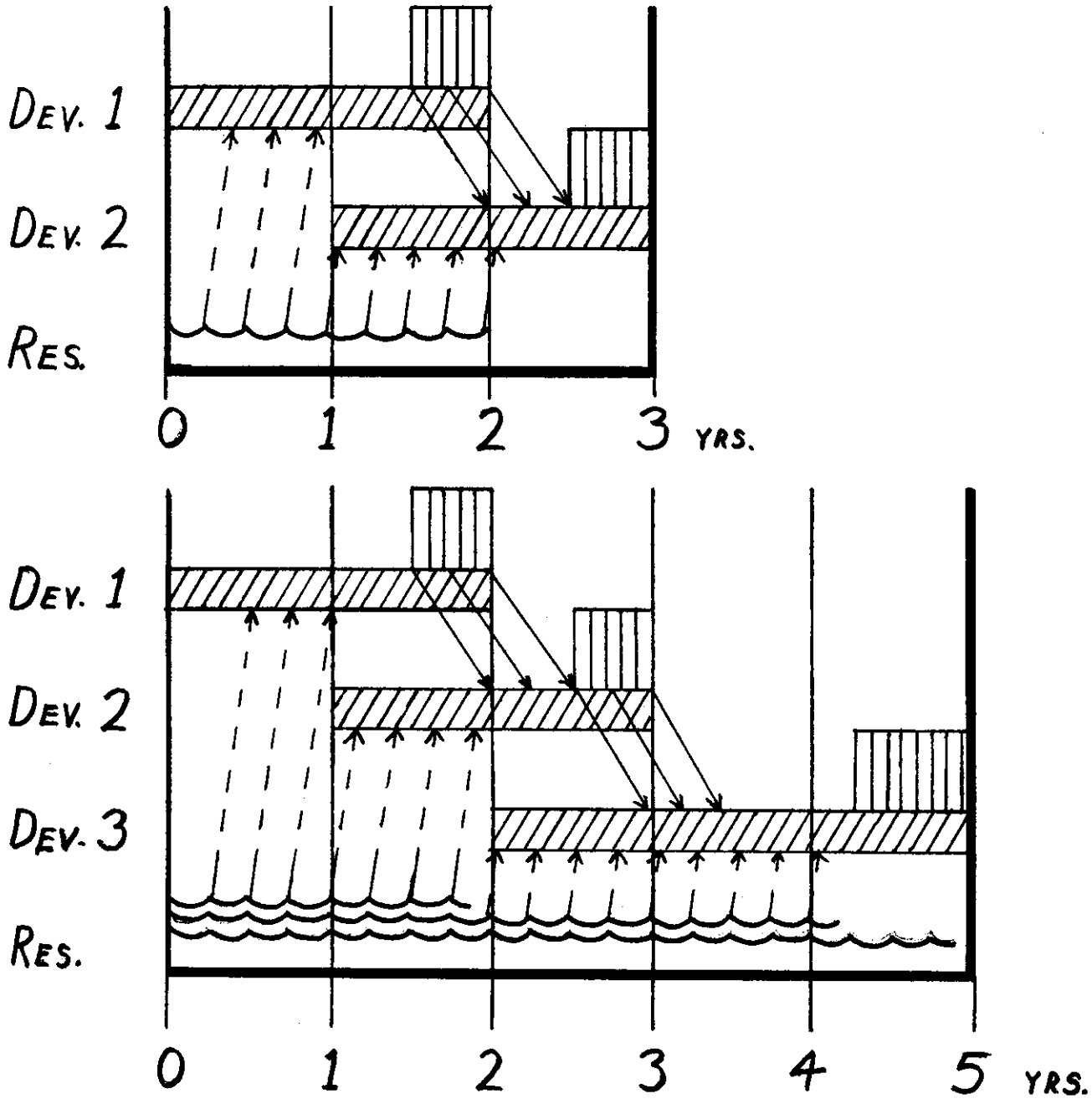


Figure 12.2: Phases of 3-year and 5-year Projects.

Single Contractor

Prime Contractor
and
Subcontractors

Circle
and
"Cooperation Expediter"

Circle of Contractors

Figure 12.3: Four Kinds of Project "Management"

also, the circle plus "cooperation expediter" schema, and my reservations about it have to do with the consideration that this undertaking, being quite difficult, will probably require more centralized planning and managing and can be accomplished without essentially total control. But perhaps one might start out on this route and then, if things did not shape up rapidly enough, shift over to the prime-contractor/subcontractor schema.

The Remaining Two Factors

Each of the project structures thus far described can be subdivided into two or more categories on the basis of the eleventh and twelfth factors listed earlier. The eleventh factor is the degree to which special arrangements of a cooperative nature are employed in the over-all project. The twelfth has to do with the involvement of research people who have been in the speech field long enough and deeply enough to be thought of as speech people rather than as computer people.

In our earlier discussions, we talked about a special support program for the collection of speech samples, for the collection and organization of specific measurements or data pertaining to speech production, analysis, and so on -- in short, a special data base to support the over-all program. We talked, also, of setting up specialized facilities in one or more locations and using those facilities via the network. Perhaps these examples suggest a dimension along which the various project structures can be scaled.

The involvement of established speech experts in a significant way in the program might be made a decision criterion. It might be made a management goal. It might be decided to be a matter of little weight. I think that I may be considered conservative on this point, for I suspect that there is, in the speech community, a large amount of expertise that the envisaged project will need if it is given the go signal. But the present purpose is not to evaluate this factor, it is only to mention it. Having mentioned it, I shall conclude.

BIBLIOGRAPHY

Alexander, A.A., Gryb, R.M., and Nast, D.W. (1960), "Capabilities of the Telephone Network for Data Transmission," Bell System Technical Journal, 39, pp. 431-476.

Allen, J. (1970), "Text-to-Speech Conversion," MIT Quarterly Progress Report, No. 99, pp. 146-148.

Andrews, F.T. and Hatch, R.W. (1970), "National Telephone Network Transmission Planning in AT&T," IEEE 1970 International Conference on Communications, San Francisco.

Atal, B.S. (1971), "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," IEEE Trans. on Computers (to be published).

Beitscher, H.R. and Webster, J.S. (1956), "Intelligibility of UHF and VHF Transmissions at 15 Representative Air Traffic Control Towers," JASA, 28, p. 561.

Bloomfield, L. (1933), Language, N.Y.: Holt, Rinehart, and Winston.

Bobrow, D.G. (1963), "Syntactic Analysis of English by Computer -- a Survey," AFIPS Conference Proceedings, Vol. 24, pp. 365-387.

Bobrow, D.G. and Fraser, J.B. (1969), "An Augmented State Transition Network Analysis Procedure," Proc. International Joint Conference on Artificial Intelligence, Washington, D.C., pp. 557-567.

Bobrow, D.G. and Klatt, D.H. (1968), "A Limited Speech Recognition System," Proc. AFIPS Fall Joint Computer Conference, Thompson, Washington, D.C., 33, pp. 305-318.

Broad, D.J. and Fertig, R.H. (1970), "Format-Frequency Trajectories in Selected CVC-Syllable Nuclei," JASA, 47, pp. 1572-1582.

Cheatham, T.E. (1966), "The Introduction of Definitional Facilities into Higher Level Programming Languages," Proc. AFIPS Fall Joint Computer Conference, pp. 613-621.

Chomsky, N. and Halle, M. (1968), The Sound Pattern on English, N.Y.: Harper and Row.

Colby, K. and Smith, D.C. (1969), "Dialogues Between Humans and an Artificial Belief System," Proc. of International Joint Conference on Artificial Intelligence, Washington, D. C., pp. 319-324.

Culler, G. J. (1969), "An Attack on the Problems of Speech Analysis and Synthesis with the Power of an On-Line System," Proc. of the International Joint Conference on Artificial Intelligence, Washington, D.C., pp. 41-48.

Cushing, S. (1969), "English as a Tone Language: the Acoustics of Primary Stress," Quarterly Progress Report of the Research Laboratory of Electronics, No. 92, MIT, pp. 351-359.

Delattre, P., Liberman, A.J. and Cooper, F.S. (1955), "Acoustic Loci and Transitional Cues for Consonants," J. Acoust. Soc. Am. 27, pp. 769-773.

Denes, P.B. and Von Keller, T.G. (1968), "Articulatory Segmentation for Automatic Recognition of Speech," Proc. 6th International Congress on Acoustics, pp. B143-B146.

Early, J. (1970), "An Efficient Context-Free Parsing Algorithm," CACM, 13, No. 2, pp. 94-102.

Flanagan, J.L. (1965), Speech Analysis, Synthesis and Perception, N.Y.: Academic Press, Inc.

Forgie, J.W. and Forgie, C.D. (1959), "Results Obtained from a Vowel Recognition Computer Program," JASA, 31, pp. 1480-1489.

Fries, C.C. (1952), Structure of English, N.Y.: Harcourt, Brace, and World.

Galler, B.A. and Perlis, A.J. (1967), "A Proposal for Definitions in Algol," CACM, 4, No. 8, pp. 204-219.

Gerstman, L.J. (1968), "Classification of Self-Normalizing Vowels," IEEE Trans, Audio and Electro-Acoustics, AU-16, pp. 78-80.

Goffman, E. (1967), Interaction Ritual, Chicago, Ill.: Aldine Publishing Co.

Gold, B. (1966), "Word Recognition Computer Program," Technical Report 456, Lincoln Laboratories, MIT, Cambridge, Mass.

Gold, B. and Rader, C. (1969), Digital Processing of Signals, N.Y.: McGraw-Hill.

Goldman-Eisler, F. (1968), Psycholinguistics, London: Academic Press.

Green, C. (1969), The Application of Theorem Proving to Question-Answering Systems, Ph.D. thesis, Stanford University, Stanford, California.

Halle, M., Hughes, G.W., and Radley, J.T. (1957), "Acoustic Properties of Stop Consonants," JASA, 29, pp. 107-116.

Hershman, R.L. and Hillix, W.A. (1965), "Data Processing in Typing: Typing Rate as a Function of Kind of Material and Amount Exposed," Human Factors, 7, pp. 483-492.

Hill, A. (1958), Introduction to Linguistic Structures, N.Y.: Harcourt Brace.

Hill, D.R. (1971), "Man-Machine Interaction Using Speech," in F.L. Alt, M. Rubinoff, and M.C. Yovits (eds.), Advances in Computers, N.Y.: Academic Press, Vol. 11, pp. 165-230.

Hockett, C. F. (1958), A Course in Modern Linguistics, N.Y.: MacMillan Co.

Hughes, G.W. and Hemdal, J.F. (1965), "Speech Analysis," Purdue Research Foundation Technical Report TR-EE65-9, Lafayette, Indiana.

Hyde, S.R. (1968), "Automatic Speech Recognition: Literature, Survey, and Discussion," Research Dept. Report No. 35, P.O. Research Dept., Dollis Hill, London, N.W. 2.

Inglis, A.H. and Tuffnell, W.L. (1951), "An Improved Telephone Set," Bell System Technical Journal, 30, pp. 239-270.

Jakobson, R., Fant, G.C.M., and Halle, M. (1963), Preliminaries to Speech Analysis: the Distinctive Features and their Correlates, Cambridge, Mass.: MIT Press.

Jones, L.V. and Wepman, J.M. (1966), A Spoken Word Count, Chicago, Ill.: Language Research Associates.

Kay, M. (1964), "A Parsing Program for Computational Grammars," RM-4283-PR, Rand Corp., Santa Monica, Calif.

Klemmer, E.T. (1969), "Grouping of Printed Digits for Manual Entry," Human Factors, 11, pp. 397-400.

Kolesnik, P.E. and Teel, K.S. (1965), "A Comparison of Three Manual Methods of Inputting Navigational Data," Human Factors, 7, pp. 451-456.

Kuno, S. (1967), "Computer Analysis of Natural Languages," in Mathematical Aspects of Computer Science, Proceedings of Symposia in Applied Mathematics, Vol. XIX, American Mathematical Society, Providence, R.I.

Kuno, S. and Oettinger, A.G. (1962), "Multiple Path Syntactic Analyzer," in Information Processing 1962, North-Holland Publishing Co., Amsterdam.

Lee, F.F. (1968), "Machine-to-Man Communication by Speech Part I: Generation of Segmental Phonemes from Text," Proc. AFIPS Spring Joint Computer Conference 1968.

Lehiste, I. (1964), "Juncture," Proc. of the 5th International Conference of the Phonetic Sciences, Munich.

Lehiste, I., ed. (1967), Readings in Acoustic Phonetics, Cambridge, Mass.: MIT Press.

Lehiste, I. (1970), Suprasegmentals, Cambridge, Mass.: MIT Press.

Lindblom, B. (1963), "Spectrographic Study of Vowel Reduction," JASA, 35, pp. 1773-1781.

Lieberman, P. (1967), Intonation, Perception and Language, Cambridge, Mass.: MIT Press.

Lindgren, N. (1965), "Machine Recognition of Human Language," IEEE Spectrum, 2, Nos. 3 and 4.

Makhoul, J.I. (1970), "Speaker-Machine Interaction in a Limited Speech Recognition System," MIT Quarterly Progress Report, No. 96, pp. 195-202.

McWhirter, N. and McWhirter, R. (1966), Guinness Book of World Records, N.Y.: Sterling Publishing Co.

Medress, M. (1969), "Computer Recognition of Single-Syllable English Words," Ph.D. Thesis, MIT, Cambridge, Mass.

Miller, G.A. (1951), Language and Communication, N.Y.: McGraw-Hill.

Miller, G.A., Heiss, G.A., and Lichten, W. (1951), "The Intelligibility of Speech as a Function of the Context of Speech Materials," J. Experimental Psychology, 41, pp. 329-335.

Minsky, M. (1969), Semantic Information Processing, Cambridge, Mass.: MIT Press.

MITRE (1964), English Preprocessor Manual, Rep. SR-132, The MITRE Corp., Bedford, Mass.

Newell, A. (1968), "On the Analysis of Human Problem Solving Protocols," in J.C. Gardin and B. Jaulin (eds.), Calcul et Formalisation dans les Sciences de L'Homme, Centre National de la Recherche Scientifique, pp. 146-185.

Newell, A. (1969), "Heuristic Programming: Ill-Structured Problems," in J.S. Aronofsky (ed.), Progress in Operations Research, Vol. 3, John Wiley and Sons, pp. 363-415.

Newell, A. and Simon, H. A. (1971), Human Problem Solving, Englewood Cliffs, N.J.: Prentice-Hall.

Öhman, SEG. (1968), "Coarticulation in VCV Utterances: Spectrographic Measurements," JASA, 39, pp. 151-168.

Otten, K.W. (1971), "Approaches to the Machine Recognition of Conversational Speech," in F.L. Alt, M. Ribinoff, and M.C. Yovits (eds.), Advances in Computers, N.Y.: Academic Press, Vol. 11, pp. 127-163.

Petrick, S.R. (1965), "A Recognition Procedure for Transformational Grammars," Ph.D. thesis, MIT, Cambridge, Mass.

Pierce, J.R. (1969), "Whither Speech Recognition?" JASA, 46, pp. 1049-1051.

Pierce, J.R. and Karlin, J.E. (1957), "Reading Rates and the Information Rate of a Human Channel," Bell System Technical Journal, 36, pp. 497-516.

Potter, R.K., Kopp, G.A., and Green, H.C. (1947), Visible Speech, N.Y.: D. Van Nostrand Co.

Quastler, H. (ed.) (1955), Information Theory in Psychology: Problems and Methods, Glencoe, Ill.: Free Press.

Quillian, M.R. (1966), "Semantic Memory," Ph.D. thesis, Carnegie Institute of Technology, Pittsburgh, Pa.

Reddy, D.R. (1967), "Computer Recognition of Connected Speech," JASA, 42, pp. 329-347.

Reddy, D.R. and Robinson, A.E. (1968), "Phoneme-to-Grapheme Translation of English," IEEE Trans., AU16: 2, pp. 240-246.

Schafer, R.W. and Rabiner, L.R. (1970), "System for Automatic Format Analysis of Voiced Speech," JASA, 47, pp. 634-648.

Simmons, R.F. (1965), "Answering English Questions by Computer: A Survey," CACM, 8, pp. 53-70.

Simmons, R.F. (1970), "Natural Language Question-Answering Systems: 1969," CACM, 13, pp. 15-30.

Siversten, E. (1961), "Segment Inventories for Speech Synthesis," based on University of Michigan Speech Research Laboratory Report No. 5.

Standish, T.A. (1969), "Some Features of PPL," ACM SIGPLAN Notices, 4.

Stevens, K.N. and Halle, M. (1962), "Speech Recognition: a Model and a Program for Research," IRE Trans. PGIT, IT-8, pp. 155-159.

Stevens, K.N. and Halle, M. (1964), "Remarks on Analysis by Synthesis and Distinctive Features," Models for the Perception of Speech and Visual Form, W. Wathen-Dun (ed.), Cambridge, Mass.: MIT Press.

Stevens, K.N. and von Bismarck, G. (1967), "A Nineteen-Channel Filter Bank Spectrum Analyzer for a Speech Recognition System," NASA Scientific Report No. 2, 1967 Contract NAS 12-138.

Stevens, K.N., House, A.S., and Paul, A.P. (1966), "Acoustical Description of Syllabic Nuclei: An Interpretation in Terms of a Dynamic Model of Articulation," JASA, 40, pp. 123-132.

Stockham, T.G. (1971), "Digital Deresonation of Acoustic Recordings," paper presented at the 81st meeting of the Acoustical Society of America, Washington, D.C.

Stockwell, R.P., Schachter, P. and Partee, B.H. (1968), Integration of Transformational Theories on English Syntax, Report ESD-TR-68-419, Vol. 2, University of California, Los Angeles, Calif.

Tappert, C.C., Dixon, N.R., Beetle, D.H., and Chapman, W.D. (1970), The Use of Dynamic Segments in the Automatic Recognition of Continuous Speech, Technical Report RADC-TR-70-22, IBM, Systems Development Division, Research Triangle Park, N.C.

Thorne, J., Bratley, P., and Dewar, H. (1968), "The Syntactic Analysis of English by Machine," Machine Intelligence 3, D. Michie (ed.), New York: American Elsevier.

Trager, G.L. and Smith, H.L. (1957), An Outline of English Structure, [Norman, Okla: Battenburg Press] Washington, D.C., American Council of Learned Societies.

Vicens, P. (1969),"Aspects of Speech Recognition by Computer," Report CS-127, Ph.D. Thesis, Computer Science Department, Stanford University.

Waterman, D.A. and Newell, A. (1971), "Protocol Analysis as a Task for Artificial Intelligence," Proc. of the Second Joint International Conference on Artificial Intelligence, London (to be published).

Watt, W. (1968), "Habitability," American Documentation, Vol. 19, No. 3.

Weizenbaum, J. (1966), "ELIZA -- A Computer Program for the Study of Natural Language Communications Between Man and Machine," CACM, 9, pp. 36-45.

Weizenbaum, J. (1967), "Contextual Understanding by Computers," CACM, 10:8, pp. 474-480.

Winograd, T. (1971), "Procedures as a Representation of Knowledge in a Computer Program for Understanding Natural Language," AI Memo, MIT.

Woods, W.A. (1970), "Transition Network Grammars for Natural Language Analysis," CACM, 13.

Woodworth, R.S. and Schlosberg, H. (1954), Experimental Psychology, N.Y.: Holt, Rinehart, and Winston.

Wolf, J.J. (1970), "Choice of Speaker Recognition Parameters," MIT Quarterly Progress Report, No. 97, pp. 125-134.

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3. REPORT TITLE
Speech-Understanding Systems: Final Report of a Study Group

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Scientific Interim

5. AUTHOR(S) *(First name, middle initial, last name)*
A. Newell, J. Barnett, J. Forgie, C. Green, D. Klatt, J. C. R. Licklider, J. Munson, R. Reddy, W. Woods

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May, 1971	147	93

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
F44620-70-C-0107	
b. PROJECT NO.	
c. A0827-5	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
61101D	
d.	

10. DISTRIBUTION STATEMENT
This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
TECH, OTHER	Air Force Office of Scientific Research (SRMA) 1400 Wilson Boulevard Arlington, Virginia 22209

13. ABSTRACT

This report provides an evaluation of the state of the art and a program for research towards the development of speech understanding systems. To assess the possibility of such systems four specific tasks were considered and evaluated. Problem areas are identified and discussed leading to the conclusions on the technical aspects of the study. A possible program for research and development is presented.

DD FORM 1 NOV 68 1473

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT