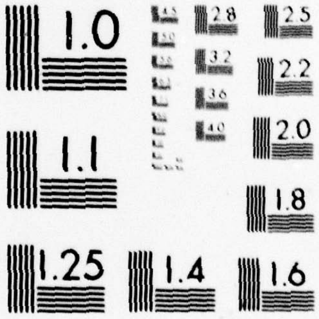
AD-A073 703 LOGICON INC SAN DIEGO CA TACTICAL AND TRAINING SYSTE--ETC F/G 17/2
VOCABULARY SPECIFICATION FOR AUTOMATIC SPEECH RECOGNITION IN AI--ETC(U)
AUG 79 J J PETERSEN, N LEE, C MEYN N00014-78-C-0692

UNCLASSIFIED NL

| OF |
AD
A073703

END
DATE
FILMED
10-79
DDC


1.0
4.5
5.0
5.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

# VOCABULARY SPECIFICATION FOR AUTOMATIC SPEECH RECOGNITION IN AIRCRAFT COCKPITS

Rohn J. Petersen

Logicon, Inc.
Tactical & Training Systems Division
4010 Sorrento Valley Boulevard
P.O. Box 80158
San Diego, California 92138

FINAL REPORT
OFFICE OF NAVAL RESEARCH
CONTRACT NUMBER N00014-78-C-0692

Approved for public release; distribution unlimited.

Submitted to

Office of Naval Research
Department of Navy
800 North Quincy Street
Arlington, Virginia 22217

August 31, 1979

This study was supported and monitored by the Office of Naval Research under contract number N00014-78-C-0692. The views and conclusions expressed in this document are those of the author alone (or of the experts pooled, whenever so indicated) and should not be interpreted as necessarily representing the official politics of the sponsors or monitoring agency, or the U.S. Government.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

**REPORT DOCUMENTATION PAGE**

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

9

4. TITLE (and Subtitle)

6

Vocabulary Specification for Automatic Speech
Recognition in Aircraft Cockpits.

5. TYPE OF REPORT & PERIOD COVERED

Final Report
Sep~~t~~ 78 ~~to~~ Jun~~e~~ 79

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

10

Rohn Petersen, Nancey Lee, Catherine Meyn,
Elaine Regelson ~~and~~ William Satzer

8. CONTRACT OR GRANT NUMBER(s)

15

N00014-78-C-0692

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Logicon, Inc., Tactical & Training Systems Div.
Post Office Box 80158
San Diego, California 92138

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

12. REPORT DATE

11

31 Aug~~ust~~ ~~19~~79

13. NUMBER OF PAGES

88

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

12 93p.

15. SECURITY CLASS. (of this report)

Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

| | |
|---|---|
| Automatic Speech Recognition | Vocabulary Specification |
| Man-Computer Interface | Human Learning |
| Human Factors | Perceptual Learning |
| Cockpit Design | Verbal Behavior |
| | F-4 Cockpit Procedures |

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The general focus of this research was to design a communication media (a vocabulary) that is advantageous to both machine recognition and human production of speech events. The problem was analyzed from a human factors perspective that centered upon the man-computer dialogue (interaction) required for cockpit application of ASR.

F-4 cockpit functions that are appropriate to ASR application were identified and candidate phrases to control these functions were defined. The

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

411171 over

1/B

phrases were categorized in terms of phrase meaningfulness, familiarity and length. The learning and utilization of the phrases were tested using a paired-associate learning task, which simulated some of the pilot's cockpit activities. In addition, the familiarity of the stimulus items was also manipulated in the experiment as a simulation of the experience level of the pilot (experienced pilots vs. novice pilots). The results indicated that phrase familiarity and stimulus familiarity had major impact on the learning and utilization of the phrases in the paired-associate task. Phrase length and meaningfulness did not appear to differentially affect either the learning or utilization of the paired associate. In addition, pretraining of stimulus familiarity did not seem to result in improved performance. Acoustic lexical confusability also was discussed in general methodological terms. The results of the study were interpreted in terms of a contextualist viewpoint with the necessity of a broader contextual manipulation being pointed out as a requirement for further research.

## TABLE OF CONTENTS

## LIST OF ILLUSTRATIONS

## LIST OF TABLES

# I. INTRODUCTION

## GENERAL BACKGROUND

The main problem facing the designers of aircraft cockpits in the near future will be to design the aircraft control functions and the other information exchange devices in such a way that the pilot and aircrew are not overwhelmed with the amount of information presented to and required from them. Current avionic systems rely primarily upon the human visual and haptic systems for information exchange with the aircraft. These systems will, in all probability, be overloaded by the information demands of future sophisticated avionics systems. The addition of another information channel would help solve the overloading problem and allow the human and the machine to interact more successfully.

The introduction of automated speech technology (AST) will eliminate a large portion of the information exchange overload by allowing the pilot and crew to perform the currently conflicting manual tasks in an efficient and expeditious manner. By allowing one of the tasks to be a verbal task, the conflict of simultaneous manual tasks is avoided. The action of a pilot controlling an aircraft is viewed as a communication between the man and the machine. In the case of sophisticated machines such as the aircraft, the rate of communication required for efficient operation is so severe that the man is hard-pressed to keep up. In the aircraft, the problem is not that the man's cognitive capacity cannot keep pace, rather the problem is that the entire dialogue must be conducted via switches, levers, dials, needles, buttons, etc. The limit is in the information that can be conveyed by the limited number of communication channels, not the information processing capabilities of the human. AST offers another communication channel to the pilot, that he is intimately familiar with and is quite natural for him to use.

The feasibility of using AST in aircraft cockpits was addressed in 1978 in an ONR sponsored study by Boeing and Logicon. One of the results of that study was a program plan for pursuing the possibility of implementing AST in the cockpit of a P-3C. The identification of considerations necessary for optimal specification of a vocabulary for automatic speech recognition (ASR) was a part of the program plan.

This project is a small but significant step in the direction of implementing AST in an aircraft cockpit. The significance of the project is primarily in the fact that it opens a new arena for speech research. Previously, the emphasis of speech research has been in developing hardware and software just to demonstrate that indeed a machine can recognize speech. That capability is now well recognized. This project expands the horizons of speech research by making the human user and the application integral parts of automatic speech technologies.

## Clutter and Information Overload in Aircraft Cockpits

The configuration of the future aircraft cockpit will potentially be cluttered, in terms of physical layout, information presentation and information demand. The impact of this clutter will be a decrement in the performance of the pilot and crew that will constitute a serious obstacle to the efficient utilization of these complex and demanding systems. The cockpit, as currently envisioned and utilized, relies on the aircrew's visual, motor and haptic systems to accomplish all man-machine interactions. In order to improve this interaction of man and machine and thereby improve the utilization of these costly systems, we must analyze the operating characteristics of each member of the communication diad.

The human side of the man-machine dialogue is the most rigid in terms of requiring specific concessions from the machine. The rigidity is primarily due to the constraints imposed by the human's cognitive apparatus. Man can process large amounts of information, provided the information is formatted in specific ways. The human is equipped with several types of perceptual systems for gathering and transmitting information, the most important of which, given the context of this paper, are the visual, auditory, articulartory, haptic and motor systems. Each system is capable of processing or transmitting information that is presented in very specific and radically different formats. The visual, haptic and motor systems are maximally attuned to information contained in the ambient light array and the tactile surface structure of the environment. In addition, all are currently being utilized in cockpit control systems in a fashion that approaches the information processing limits of each system. These systems are spatially specific systems, that is, the receptive organ and transmitive masculature must be oriented to a specific location in order to extract or transmit information. It is this limit in processing capacity that presents the system designer with his problem. As avionic systems become more complex, more information is made available to and required of the aircrew. Theoretically, this information contributes to the increased efficiency with which the aircraft and crew can accomplish their mission. Given that the visual, haptic and motor systems are verging on being overburdened, the designer is faced with the dilemma of either ignoring some of these desirable capabilities affored by the advanced avionic technologies or addding an additional crew member to handle the burden. The latter alternative is not usually a viable option because the aircraft is spatially constrained and the expansion of that space is a costly proposal. The former alternative is also unattractive because the success of the new aircraft is usually predicated on the utilization of the most advanced technologies available to provide a tactical advantage in a particular area.

## Automated Speech Technology, A Potential Solution

Automated speech technologies provide a means by which the aircrew's auditory perceptual system and speech generation system can be incorporated into the advanced cockpit environment. AST can be an effective substitute for the motor responses (key/ button pressing) currently used to input information in avionic systems. Eliminating the motor response eliminates the need for some of the switches or buttons and thereby increases the available display space in the cockpit. In addition, speech generation by the machine will constitute another form of information output for the avionic system, in essence providing another display to the crew.

The most important aspect of AST utilization is the elimination of the information overload of the visual, haptic and motor systems. The incorporation of the human auditory perceptual system and speech generation system into the information processing system does not automatically decrease the workload on the overall cognitive system. The human cognitive system is basically a single channel information processing device. The single channel limitation is thought to be imposed by the attentional system. However, the advantage of adding auditory processing to the information exchange formats is that audition is not spatially constrainted and therefore can operate independently, without interfering with ongoing visual, haptic or motor exchanges.

Human perceptual systems have evolved a mechanism to overcome the constraints of the single channel attention system. The primary apparatus utilized is the process of perceptual learning. Through perceptual learning the perceptual systems attain the capability of automatically processing certain types of information and essentially bypassing the bottleneck created by the single channel attentional mechanism. Human cognition then can be thought of as both a serial and parallel processor. The visual system as a perceptual system can be easily overloaded. This overloading usually takes the form of inputs that are spatially located outside the prime visual field. The same is true of the motor and haptic systems. Given the spatial constraints of these systems the addition of auditory exchange will expand the information field of the operator, eliminating the spatial constraints on information exchange. In general, the utilization of the auditory exchange expands the information processing capacity of the cognitive system and serves to distribute the perceptual processing load, thereby eliminating overloading of a single perceptual system.

## The Role of Human Factors in AST Applications

AST is for many applications a significant human factors development. The natural language potential of an AST system offers a unique interface with a machine. However, the incorporation of AST does not guarantee that the human will be better served by the system. The design restrictions of the AST system could conceivably make the system more difficult to use than a well factored conventional system. The point is that the implementation of an AST system requires a well human-factored design, just as any other man-machine interaction does.

Some of the important human factors considerations for AST design are listed below. This modicum of human factors considerations includes the major characteristics of a well designed man-machine dialogue. The main point of identifying these characteristics is to design a system that is readily compatible with the user's cognitive process and structures. User acceptance is the primary criterion of a well human-factored system. If the user likes the machine and can easily interact with it, he will use it in an efficient and creative manner. If, on the other hand, the system is cumbersome or in some other way difficult for him to use, he will avoid using it. If the designer will attempt to incorporate as many as possible of these dialogue characteristics, the resultant system will find a high degree of favor and acceptance among users.

- Speaker Independence - loosely defined as the ability of the AST system to recognize the speech of all users without an elaborate training session with the machine. Extensive training periods are not only time consumming and costly, they are equally quite fragile. That is, should the speaker's voice change (from a cold, oxygen, or excessive acceleration forces) the machine will have to be retrained.
- Vocabulary Design - the design of a vocabulary that fits easily within the cognitive structure of the user requiring little or no rote learning and, at the same time, a vocabulary that facilitates machine recognition by eliminating auditory similarity between vocabulary items.
- Human Channel Capacity - the user can typically process only 5±2 items at a given time. Therefore, the dialogue should not be arranged in a manner that requires the user to store more than 5±2 items at a given time. The size of the item is dependent upon the verbal (learning) ability of the user.
- Machine Response Time - machines that respond too fast or too slow in a natural language dialogue detract from the interaction by creating unnatural breaks in the dialogue or by eliminating breaks where they should be

imposed. This detracts from the user's normal cognitive processing making the system more difficult for him to use.

- Continuous Speech Recognition - allows the user to speak naturally without the artificial pause between words or phrases that is required by the individual word or phrase recognitions systems. The speaker does not have to learn to speak to a continuous speech recognition machine.
- Voice Recognition Feedback - when the user makes a verbal response the system must respond in some way to indicate that it has recognized the verbal response. The machine's acknowledgment of the verbal response indicates to the user that he has voiced the response in a manner that the machine can recognize. Typically the user finds this form of feedback quite reinforcing. If no acknowledgment is forthcoming from the machine, the user immediately knows he has misvoiced the response and must try again. The procedure is very similar to the shaping procedures of operant conditioning and to the focus of the role of information feedback in motor learning.
- Application Considerations - AST, like most new technologies, is very impressive and fun to play with, think about, etc. This novelty and enthusiasum can be misdirected. It is very tempting to design a system that makes maximal use of AST, even when the application does not call for it. Judicious use of AST may be the best course, at least for now.

## DESCRIPTION OF THE SPECIFIC PROBLEM

This project addresses two related problems: first, the determination of some of the major factors that influence the selection of a vocabulary for automated speech recognition (ASR), and second, the development of a suitable vocabulary for selected F-4 cockpit functions on the basis of the factors identified in this study.

### Project Orientation

As pointed out above, this project is oriented toward the human factors aspects of implementing ASR. The human factors aspects of ASR revolve around two main components, the human's (user's) mental and physical abilities and the specification of a device that takes advantage of these abilities. For our purposes, the human's abilities to learn and use vocabularies are of interest. A device that takes advantage of these abilities in the ASR context is the speech recognition hardware and software. While we are not directly concerned with evaluating hardware or software per se, we are interested in the user's interaction with the hardware and the software. There are a large number of capabilities of interest including continuous speech recognition, speaker independence, etc. However, the state of these arts is still evolving rapidly and hence meaningful human factors studies would be premature. The ability to recognize human speech on a phrase-by-phrase basis (individual word recognition, IWR) is well established, however, and so we are in a position to address the vocabulary definition problems.

Our concern here is the grammatical constructs that can be employed to facilitate the dialogue between the machine and the man. Grammar is an abstract construct that structures the dialogue helping each member of the diad extract meaning from the communication. There are several aspects to grammar, and each of these aspects has an effect on the potential man-machine interaction. Grammar can be divided conveniently into several components, including syntax, morphology, phonology and something less rigorously defined called the lexicon. Syntax is a rule system for determining the structure of an utterance. It is a set of rules for determining what-can-precede-or-follow-what in the sentence, the who-can-do-what-to-whom rules of communication. Morphology has to do with the structure or composition of individual words, and phonology is concerned with the way words or phrases are pronounced. Each of these are rule systems that can be manipulated or used to derive utterances. The lexicon, on the other hand, is an arbitrary system that encompasses the meaning or semantic relationship between the language (words) and the real world which the words reference.

In order to systematically determine the effects of grammar on the man-machine dialogue, each of these grammatical components must be addressed separately and together, with the effects on the dialogue being noted. It is exactly this problem that this project addresses. However, a consideration of grammer in its entirety is too large a proposition for a single investigation.

## Project Scope

In order to limit the scope of this project to something more manageable, our studies are limited to an investigation of some of the interesting aspects of morphology, phonology and the lexicon. Syntax will not be explicitly addressed during this investigation because the implementation and study of more than one syntax is a significant effort. Given that the software to control the experiments is also a significant undertaking, the prudent course is to investigate questions that can be feasibly and meaningfully addressed within the resources of the project.

Once the experimental software has been developed, the syntax question can be efficiently addressed in a later investigation with a reasonable software effort. For this reason syntactical considerations are not part of the current investigation. The phonological and lexical aspects of the human-machine dialogue grammar are the primary focus of this project. However, the importance of syntax should not be underrated.

## REPORT OVERVIEW

The approach taken to the resolution of the problems stated above was to subjectively identify candidate vocabulary considerations and then empirically test the influence of several of these factors on a human-computer dialogue. Below is a short section-by-section overview of the report.

### Technical Review

The concept of a human-computer dialogue is analyzed from the perspectives of user (human) considerations and machine considerations. This concept is further explored in the context of applications of AST as a formal dialogue device. The final topic of discussion for this section is an explicit statement of project objectives and an overview of the methodology that was used to accomplish them.

### General Methodology

The methodology utilized to accomplish the project objectives are presented in a more or less chronological order. The first step was to establish a direction for the project, based upon the statement of project objectives and the informed recommendations of Logicon's speech group. This effort was followed by a formal statement of the experimental design and an effort directed at identifying F-4 cockpit control functions that are amenable to ASR constraints.

## Experimental Methodology and Results

The specific methodology used to collect the learning and utilization data are described in detail, and the analyses of the results are reported for the two experiments that were conducted.

## Conclusions and Recommendations

The data from the experiments are discussed, and conclusions relating to the selection of vocabulary items are drawn. In addition, the problems of acoustic confusability and the general problem of vocabulary selection given multiple criterions are discussed. Finally, recommendations for further inquiry are made and discussed.

# II. TECHNICAL REVIEW

In this section the general concept of human interaction with computers is discussed as it relates to the vocabulary specification problem. The first portion of the section addresses the issue of human-computer dialogues in general. The second portion deals with the issues raised when the medium for the dialogue is voice and hence automatic speech recognition. The last portion of the section establishes the specific objectives for the project and introduces the methodology that was used to accomplish the objectives.

## HUMAN-COMPUTER DIALOGUES

In order to identify vocabulary attributes that affect the man-computer interaction, we must first analyze the concept of a man-computer dialogue. This must be done in order to determine which factors are specific to vocabularies and which are specific only to dialogues in general.

Martin (1972) has considered this problem in detail and has identified three major levels of consideration. The first is the functional level, that is, identifying which functions should be performed by the machine and which by man. The second level is a procedural level that involves determining the functions of the system and what procedures should be employed to accomplish them in an efficient and error-free manner. The third level of consideration involves a format level (Martin classified these as syntactical considerations, but to avoid confusion with syntax in the grammatical sense, we are using the term format considerations) which determines the exact form of the interaction between man and machine. In keeping with Martin's analysis, we will address each of these three levels of consideration separately.

## Functional Considerations

Functional considerations involve recognition of the divergent capabilities and disabilities of the human and the computer. Based upon this analysis the systems designer must design a system that takes advantage of as many capabilities as possible while avoiding the limitations imposed by the disabilities.

Theoretically the perfectly designed system is attainable; however, in practice, like most idealizations, it is seldom if ever achieved. A reasonable approximation is, however, well within the grasp of most thoughtful designers.

The major positive attributes of human information processing include the ability to:

a. extract invariants from unfamiliar situations that are the same or similar to those experienced under different circumstances.
b. handle problems that require a degree of judgment to arrive at a solution.
c. deal with the ambiguities.
d. process information in a seemingly parallel manner.

These attributes must be taken advantage of in designing the system. The human processor also has some attendant negative attributes, such as:

a. the limited capability of the attentional system in terms of the amount of information it can process at a given time.

b. the limited storage capacity of short term memory.
c. the context-oriented processing control system.
d. the limit in the processing speed.

The negative attributes are obstacles that the designer must overcome in designing his system, and if he does it cleverly, he can use these attributes to the benefit of the dialogue. The computer offers the designer a means of overcoming many of the problems, merely by taking advantage of the computer's positive processing attributes, such as:

a. speed of processing
b. accuracy of processing
c. capability for large information processing
d. ability to attend to minor details in a comprehensive manner
e. tolerance for redundancy
f. ability to monitor a large number of inputs seemingly simultaneously

## Procedural Considerations

Procedural Considerations involve determining which procedures will ultimately lead to the accomplishment of the function. Procedural considerations include factors such as:

a. Who will use the system?
b. Will help be needed?
c. How will errors be handled?
d. What are the locations and types of input/output terminals?
e. What is the level of training of the user?
f. What type of dialogue is to be used?
g. Can dialogue shortcuts be incorporated?

h. What other activities will the user be engaged in?
i. Where and how does he receive information for dialogue participation?

Some of these considerations are meaningful for our study; others are treated as given in the application and hence require little discussion.

## Format Considerations

Format considerations are concerned with the actual input and output messages that are incorporated in the dialogue. This is the level of consideration that we are primarily concerned with in this study. Vocabulary is a major format consideration, since the vocabulary is the vehicle of information exchange. There are several other format considerations that may be addressed, including:

a. the amount of information transmitted per dialogue message.
b. the feedback mechanisms for both the man and the computer.
c. the procedure for correcting errors.
d. the question of recognizing misrecognitions and compensating for them.
e. the coding of instructions to the user so that he will recognize them as instructions.
f. the time the system takes to respond to a given input.

While each of these considerations is important, they will be treated during the study in the same way as the procedural considerations: namely, they will remain constant across all experimental conditions. The format considerations of more specific concern here are those factors that may be considered characteristic of the vocabulary.

Vocabulary has a definite effect on both members of the man-computer diad. The human can be constrainted or aided by the vocabulary. The two major cognitive processes that interact with vocabulary are verbal learning and verbal utilization, and each of these require separate as well as joint consideration. The computer also can be constrained or aided by the vocabulary, both in terms of learning to recognize the speech patterns of a particular user and in the general facilitation of the recognition process.

## AST AND THE MAN-COMPUTER DIALOGUE

A well designed man-computer dialogue is built on a foundation provided by the vocabulary of the dialogue. In that respect, the vocabulary is a major limiting factor in determining the utilization and efficiency of the dialogue. Dialogue in this sense implies the passing of information back and forth between man and machine, each providing closure of the feedback loop of the other. In this way the behavior of each member of the diad is controlled or structured by the dialogue. When man and machine must interact to accomplish a specific mission, the design of the dialogue will have a profound effect on the success of the mission. There are several aspects to the design of the dialogue foundation, the vocabulary, that must be considered.

The optimization of a vocabulary is important from the standpoint of the user's acceptance of the system and from the standpoint of optimizing the recognition accuracy of the system. From the user's point of view, the vocabulary that he is allowed or required to use in communicating with the machine can have profound influences upon the success of the interaction. If the vocabulary is very natural, (in terms of both the mode of communication, i.e., buttons, dials, footpedals, keyboards, touchpanels, speech, etc., and in terms of simple and meaningful responses) and has a high degree of association to the actions or requests they symbolize, the vocabulary will fit well within the user's current cognitive context, and the user will readily accept the system. If the vocabulary is only marginally related to the actions and requests the items are designed to represent, and if the vocabulary items do not provide a good approximation to natural language, the user will be less inclined to use the system because it will require an extreme modification (through learning) of his cognitive structures and processes.

AST offers the user a very natural mode of communication which can be structured in very simple and meaningful words or phrases which do have a high degree of association to the actions or requests they symbolize. An AST mediated dialogue is an extremely attractive device for man-computer communications. The structure of the dialogue is provided by the vocabulary and hence, the grammar of the dialogue. In the paragraphs that follow we will discuss the relation of man with the vocabulary of a dialogue and the relation of the machine to the vocabulary. In addition we will review selected applications of ASR in operational environments.

## ASR Vocabularies and Human Interaction

When considering human interaction with vocabulary there are two processes that require our attention. They are the process of learning the vocabulary and the process of utilizing the vocabulary once it has been learned. These processes are independent, to some degree, and we should expect that some characteristics of the vocabulary would differentially affect them. The learning process is an instance of verbal learning. Verbal learning has a long history of investigation and through a review of this literature the relevant learning variables were identified. The utilization of vocabularies and the interaction of learning and utilization are also matters that must be considered when defining a vocabulary for an ASR mediated man-computer dialogue.

Utilization of a vocabulary is dependent upon two independent aspects of the dialogue, namely, the perceptual and cognitive characteristics of both members of the diad. The human will impose certain constraints on the dialogue, as will the computer. It is these constraints that need to be identified for vocabulary development and hence effective dialogue design. The factors that appear to provide the most promise of being significant human constraints include the dimensions of verbal characteristics delineated by Hall (1971) and a number of factors that are specific to the context of AST dialogue vocabulary development. Hall has identified a number of dimensions of verbal characteristics that have dominated verbal learning for a number of years. However, to investigate all of them relative to our particular application would be a herculean task. Within the constraints of this effort the prudent action is to identify the factors that seem most relevant and investigate their effects.

Factors that are specific to the context of AST dialog vocabulary development include:

a. the length of the vocabulary item
b. the total number of items in the vocabulary
c. the redundancy within vocabulary
d. the relative independence of each vocabulary item in terms of semantic similarity
e. the format of the vocabulary sequences

Learning and utilization may interact in a manner that precludes optimization of the vocabulary along either dimension. The optimization of a vocabulary for learning may preclude optimization for utilization. The same may be true of the factors identified as constraints in the discussion above. In addition, it may be that minor concessions on the part of most of the constraints may be all that is needed to allow reasonable definition of a good vocabulary.

ASR Vocabularies and Machine Interaction

The computer demands certain concessions in vocabulary design when applying AST to cockpit functions. The computer is limited, in terms of hardware and software, by the technological state-of-the-art. Several constraints or unknowns have been identified and proposed as areas of research in other documents (Feuge and Geer, 1978). Briefly, the areas of interest include:

a. continuous speech recognition
b. speaker independence
c. microphone use
d. acoustic constraints on recognition

Each of these areas requires empirical investigation and in some cases, such as continuous speech recognition, require considerable methodological and theoretical work.

Within the context of hardware and software constraints on vocabulary definition, the investigation of acoustic confusability is the primary candidate for investigation. To demonstrate how acoustic confusability might arise, it is instructive to consider a typical automatic speech understanding system, consisting of a voice input preprocessor, a mini-CPU with the usual peripherals, a software recognition algorithm, and some reference data.

The voice input preprocessor takes the analog time domain signal from a microphone, extracts spectral and time information, digitizes it and periodically (every few milliseconds) sends this information to the CPU for further processing.

The system now initiates a time normalization process. That is, regardless of the length of the phrase, all of the data in the input array are squeezed into a data structure which has a fixed number of time slots. This is done by dividing the entire input array into the number of segments. If a feature is set for a quarter or more of the samples in each segment, that feature will be set in the corresponding time slot of the feature array.

The input feature array is then compared on a bit-by-bit basis with previously established reference arrays which describe the talker's speech patterns for all items in the vocabulary. The reference array producing the highest correlation is selected as the phrase which was spoken.

The problem with the above selection process is that all incoming feature arrays will produce correlation scores when compared to the reference arrays. If the system simply chose the highest score after comparison, every input would produce a recognition whether valid or not. The use of a minimum score threshold value provides some rejection capability but introduces the possibility of non-recognition of a valid input.

Another source of confusion lies in required data reduction which is necessary for reasonable speed on a man-computer system. Because of the significant reduction that occurs in the voice input preprocessor, two acoustically distinct signals may appear quite similar on output to the CPU. In the CPU the first processing which takes place is a time normalization. Though necessary, much acoustic information is lost during normalization.

Vocabulary design can also contribute to the accuracy of the machine in recognizing speech and the accuracy of the human in using the various vocabulary items. When the improvement of a vocabulary is being considered, the exclusion of acoustically similar words is a factor that should contribute markedly to the adequacy of the dialogue. Vocabularies that include a number of acoustically similar items will suffer from poor recall on the part of the user, poor recognition by the machine and will be a continuing source of aggravation. A well defined and acoustically unambiguous vocabulary, on the other hand, will contain few acoustically similar items with the resultant elimination of most recognition errors. Examples of the effects of acoustical similarity on the processing of phonemic strings by humans can be seen in the work of Tikofsky and McInish (1968) and Conrad (1963, 1972) with individual words, Brown and Hildum (1956) with triple-phonem syllabus and Conrad (1964) with individual letters. Conrad (1963), for instance, found that confusable words such as cat, rat, bat, hat, etc., are not recalled as well as nonconfusable words, such as fish, grid, lens, spoon, etc. Similar results were demonstrated in the other studies indicating that acoustic confusion may be a major factor in constraining human utilization of verbal materials.

Artificial systems were designed to simulate man's verbal abilities and therefore might be expected to suffer from the similar difficulties. This expectation is verified to a degree by attempts to implement ASR systems in real world applications (Grady, Hicklin and Porter, 1977 and Grady and Hicklin, 1976). Grady and Hicklin (1976) describe a metricization of acoustic similarity as it pertains to computer speech recognition. The viability of their technique was also demonstrated in the context of a training system for ground controlled approach controllers.

## ASR and Operational Environments

In reviewing the literature we came across a study that dealt with the topic of implementing ASR in an aircraft cockpit. This study was done as a master's thesis at the Naval Postgraduate School in 1977 by Anthony Quartano. It dealt with vocabulary type manipulations for ASR implementation in an aircraft cockpit. Quartano was interested in the potential of implementing ASR in aircraft cockpits, the P-3C in particular. The two problems that Quartano stated as the focus of his research were (1) "What are the best words to use, as far as the human operator is concerned, and how will this command vocabulary be developed?" and (2) "(Will) the newly developed vocabulary be compatible with the constraints of the voice recognition machine?" Actually, this study only addressed two very specific problems concerning vocabulary specification, and unfortunately the inferences that may be drawn from the study do not go very far toward answering the global issues Quartano set up as the problems his study was addressing. Nevertheless, his study is enlightening as it provides information about two of the morphological and syntactical aspects of vocabulary design. Quartano found that his subjects (an aviator group and a non-aviator group) preferred to use (that is, generated as descriptions) two-word phrases as opposed to one-word phrases; however, no performance increment could be demonstrated as attributable to the two word commands. The results also indicated that the aviators tended to generate what was called "Descriptive Phrases" rather than "Command Phrases" while non-aviators did not exhibit any such tendencies. An interesting observation was made to account for this finding. Three out of the five (60 percent) P-3C keyset functions used in the experiment are currently described by "descriptive phrases," and the other two are "command phrases." The data corresponds quite closely, in that 62 percent of the phrases generated were descriptive phrases. This correspondence is perhaps indicative that the task used in this experiment was not suited to the kinds of questions that were being asked. The data relating a preference for two-word over one-word phrases is interesting and should be investigated further. Quartano's study was similar to this project in that the intent of both studies appears to be similar; however, the methodologies for addressing the questions are quite dissimilar.

Several investigators have reported or have planned studies which examine the effects of a number of cockpit factors on the accuracy of ASR (Coler, Plummer, Haft and Hitchcock, 1977, Curran, 1977, Montague, 1977). These factors included the use of an oxygen mask by the speaker, G forces, vibration, cockpit temperature, extraneous noise and mission duration. The major findings of these studies were:

1. Voice quality degrades after 0.5 hours with an oxygen mask (Curran, 1977).
2. Voice quality degrades under high (4.3 g) vibration (Curran, 1977).
3. Voice quality degrades under higher levels of g (Curran, 1977, Montague 1977).

The dialogue parameters of words said and vocabulary subset size (Coler, et al., 1977) have also been manipulated. Coler et al. (1977) report that as expected, recognition accuracy decreases as the vocabulary subset size increases.

## PROJECT OBJECTIVES

In the previous sections an overview of the problem was presented as well as rationale for limiting the scope of this project to the man-computer dialogue aspects of ASR. In addition, the scope of this particular investigation was delineated as an investigation of the interaction of the human's ability to learn and use a vocabulary as a function of various aspects of the morphological, phonetical and lexical characteristics of the vocabulary. In keeping with this orientation, the following specific objectives are delineated.

1. Specify general, subjective considerations (problems, variables) that may be appropriate to the specification of a vocabulary for an operational ASR application.
2. Based upon the considerations that result from the accomplishment of Objective 1, identify considerations where the solution of the consideration is not intuitively obvious.
3. Empirically test these considerations in order to determine a good solution.
4. Based upon the considerations and solutions, define a candidate "good" vocabulary for implementing ASR to control specific functions in an F-4 cockpit.

The general methodology that was used to accomplish these objectives is outlined below and expanded upon in the next section.

## Identification of Candidate Vocabulary Considerations

The first step in conducting this study was identification of candidate vocabulary considerations. This involved a review of the pertinent literature and the inputs from Logicon's speech recognition group. The activity resulted in the generation of a number of potential considerations for ASR vocabulary development.

## Design of the Experiment

The design of an experiment is a systematic attempt to arrange conditions so that comparisons between a subject's performance within those conditions will be affected only by a single factor. This allows the experimenter to conclude that a particular factor is indeed responsible for the change in performance. As is usually the case with controlled experiments, the experimental situation is actually just a controlled simulation of some "real life" situation. For this study the simulation is of a hypothetical cockpit environment in which the pilot is using ASR control functions to perform some of the operations he would normally do manually. The situation is such that an event or series of events occur that make it necessary for the pilot to perform a particular action. This action in this context would be to utter a particular phrase that causes the aircraft or a component of the aircraft to act in a specific way. In the laboratory we simulated this cockpit situation by presenting the subject with a particular stimulus, in this case a string of letters, and requiring him to respond with the appropriate utterance. The situation is analogous to the cockpit situation in that the subject is responding to a stimulus in the environment just as the pilot would be responding to an environmental situation.

In the real world there are other factors, besides stimuli from the immediate environment, which influence the accuracy and timing of such a response. One of the most important of these factors is the experience of the pilot. That is, how well can the pilot recognize the significance of the environmental situation, and can he associate that situation with the proper response or utterance in our situation? We can simulate the experience of the pilot in the laboratory by controlling the amount of perceptual and associative learning the subject is allowed to engage in. Perceptual learning refers to how familiar the subject is with the stimulus, while associative learning refers to how well the subject can link a specific stimulus to a specific response. By designing an experiment that varies both the characteristics of the vocabulary and the "experience" level of the subject (pilot) we can observe the interaction of these manipulations that would escape our scrutiny if treated separately.

## Selection of ASR Appropriate F-4 Cockpit Functions

The selection of ASR appropriate F-4 cockpit functions is a task that involves developing a rationale for determining the appropriateness of ASR control for a particular aircraft function. The rationale and the candidate systems are listed in the next section. This step was necessary because of the orientation of the project in providing a usable vocabulary designed for actual implementation in an F-4 cockpit.

The specifics of the experimental methodology and the results of the experiments are presented in Section IV; Section V is a discussion of the experimental data, the problems that were encountered and recommendations for future efforts.

# III. GENERAL METHODOLOGY

This section is concerned with the specification of the general methodological efforts that were undertaken prior to conducting the formal experiments. These efforts include the delineation of a set of subjective assessments of potential questions that can and should be addressed experimentally. Following this effort a more formal specification of the experimental design and procedures is given. The last general discussion of this section has to do with the determination of F-4 cockpit functions appropriate to ASR control.

## IDENTIFICATION OF POTENTIAL VOCABULARY CONSIDERATIONS

One of the initial project efforts was to develop a set of subjectively important considerations for specifying vocabulary items. These considerations were generated by members of Logicon's technical staff who are experienced in applying ASR. They were asked to respond to the question: "If you had to define a vocabulary for a specific application, aside from knowledge of the application, what factors would influence how you chose the vocabulary items?" Although the most frequently mentioned consideration was the syntax being used by the ASR algorithm, there was some excellent discussion of such topics as vocabulary familiarity, word and phrase length, acoustic distinctiveness, vocabulary size, phrase meaningfulness and stimulus familiarity. The result of this meeting was a fairly substantial list of potential considerations with a rationale associated with most of them. Listed below are the higher priority considerations and a rationale for subjecting, or not subjecting, the consideration to empirical scrutiny.

a. Phrase Length - Phrases of several syllables in length are more easily recognized by the machine than very short phrases. One syllable words such as "eight" have a greater tendency to get lost by the machine. Longer phrases, however, may require more cognitive work on the part of the user; therefore the user may prefer to use, and perform better using, short phrases. This consideration leaves the designer caught in the middle, not knowing which direction to turn, an excellent candidate for empirical investigation.

b. Phrase Familiarity - The problem with phrase familiarity is not that it pulls the designer between two extremes, rather it tends to leave the designer in somewhat of a quandry. What does familiarity really mean? The primary problem is that there is no obvious metric on which to base a decision about familiarity.

Is familiarity best measured as an objective count of the word or phrase in the literature, or would a subjective evaluation of the meaningfulness of a word be closer to what we are actually referring to as familiarity? Perhaps the best answer is an empirical one that relates directly to our context of automatic speech recognition.

c. Acoustic Distinctiveness - Words which sound different to the human ear may be confused by the machine. An example is the "five"/"nine" confusion, which seems to plague several speech systems and to be independent of the particular recognition algorithm utilized. The opposite type confusion is also found where the machine can recognize the utterances beautifully, however the human/user tends to confuse them either in memory or in production. It is possible that some general rules will result from an investigation of this problem.

d. Learning the Vocabulary - One vocabulary item may be easier to learn than another; however, there is no a priori way of determining that, although it is likely that the variables described above will influence the rate of acquisition. Another factor also must be taken into account when considering learning, and that factor is user experience (that is, the amount of previous learning in which the subject has engaged in the context of his current learning task). Perhaps we could better state the problem as taking a look at the amount of learning the subject must accomplish within a given period of time. The experienced subject is already familiar with certain aspects of the environment, while the novice must assimilate information from a broader spectrum of the environment and hence cannot keep pace with the experienced person focusing on the new portion of that environment. A likely hypothesis, but as yet it has not been demonstrated in the area of perceptual learning.

e. Vocabulary Size - The smaller the vocabulary, the better machine recognition tends to be. With regard to the use of the vocabulary, however, there may be a tradeoff between vocabulary size and degree of learning. This variable has the potential of being critical to the functionality of ASR.

f. Referent - The use of words or phrases as control devices in a cockpit implies that the word or phrase has a physical referent in the real world. In the cockpit environment, the referent may be of two varieties: the actual function response desired or the currently available control device.

Most of the other considerations that were voiced either supported one of the considerations above, was ranked as a minor factor, or was more closely related to a syntactical consideration. A note of caution that was clearly articulated by this group was that the particular syntax and the particular ASR system upon which the experiments are conducted might impact the results of the study. This seems particularly evident upon the questions relating to the machine recognition abilities. However, it should be recognized that the human's performance also may be influenced by these factors, but this influence will be much harder to determine.

## SELECTION OF VARIABLES IN EXPERIMENTAL DESIGN

The basic premise of an experiment is that all important factors are systematically controlled so that the variation of one factor from condition A to condition B can be confidently said to be the sole reason for any change in the subject's performance. For the most part, the laboratory is the only place where this type of rigid control is possible. For example, a subject (or group of subjects) is given phrases that are long in one instance and short in another. The point of the study is to examine the subject's performance under both conditions and determine if the difference is significant enough and reliable enough to warrant concluding that long (or short) phrases cause the subject to verbalize the phrase more accurately or more expeditiously. The performance that is measured need not be limited to human performance. The machine performance can be treated in an analogous manner.

This study is primarily concerned with two aspects of human performance, the initial learning of a vocabulary and the efficiency of utilizing the vocabulary. This point is to monitor user performance for changes that may be attributed to a particular variable or interaction of variables, that is, variables acting simultaneously to produce change in performance. The variables used in the experiment and a brief explanation of the variable are presented below. The variables are addressed in the experiments reported in Section IV.

a. Phrase Familiarity - Phrase Familiarity is a variable that is easily scaled by soliciting the subjective assessments of a user group. For the purposes of this study two levels of familiarity were investigated, highly familiar phrases and unfamiliar phrases. Theoretically, familiarity is thought to be related to the number of times the phrase has been encountered in the past. Logically, it seems the more familiar the phrase the easier it should be for the subject to retrieve the phrase from memory and learn the association between the phrase and the stimulus. However, there may be a cost associated with highly familiar phrases due to increased proactive inhibition, that is, inhibition from previous associations of the phrase with other stimuli. So the real value of phrase familiarity for our application must be determined empirically.

b. Phrase Meaningfulness - Like familiarity, meaningfulness is a qualitative type variable that requires a subjective rating of each phrase in order to scale the relative meaningfulness of the phrases in the vocabulary. In theoretical terms meaningfulness can be thought of as the number and strength of the memorial associations related to the particular phrase. The more associations, the more meaningful the phrase. In addition, the more concrete (real worldish) the association, the stronger the strength of the association and hence the more meaningful. Again, it appears from a logical point of view that the more meaningful the phrase is, the easier it should be for the subject to recall the phrase from memory and learn the association between the phrase and the stimulus. However, again, there also could be the proactive inhibition problem where the number and strength of the previous associations interfere with the learning and utilization of a new association.

c. Stimulus Familiarity - The determination of whether or not being familiar with the stimulus (environment) facilitates the learning or utilization of a paired-associate (stimulus and response phrase) may be an important consideration for determining when to introduce an ASR vocabulary as a control function. Two levels of familiarity are appropriate to this variable in our experimental tasks, familiar stimuli (words) and unfamiliar stimuli (nonwords).

The design of our experiment is a within-subjects type design. This type of design, coupled with laboratory control of extraneous variables, permits unambiguous conclusions to be drawn from the results of the experiment. Within subject variables are variables that can be manipulated in such a way that each subject experiences all levels of a particular variable. The advantage of this type of design is that all comparisons within variable levels and between variables are based upon the data from each subject in the experiment. This eliminates the possibility of major individual differences effects contributing as sources of variance in any of the differences that may be found. In addition, it significantly reduces the number of subjects that are required to complete the design.

The control of the various confounding factors was also a major consideration in the experimental design. Most of the confounding type variables have been identified previously. A variety of control techniques, including balancing, constancy of conditions, elimination, and counter-balancing were used to control these variables.

There are several other variables that may be important in vocabulary specification. Three of these additional variables are described below and taken up in the later portion of this report.

a. Phrase Length - Phrase length can be measured in terms of the number of syllables in the phrase. Two levels of this variable were considered, long and short. The tradeoff in phrase length may be that long phrases are easier for the machine to recognize, but difficult for the man to remember; short phrases are easier for the man to remember, but more difficult for the machine to recognize.

b. Acoustic Confusions - There are basically two perspectives that must be addressed when considering acoustic confusability, confusions by the machine and confusions by the user. Both of these aspects must be scaled prior to the experiment. The scaling of human confusions can be addressed in at least two different ways, subjective ratings and overt performance confusions. Other researchers have demonstrated that acoustic confusions are manifest for human subjects during short term memory tasks. However, it should be noted that this sort of overt measure of performance is very time consuming and costly.

c. Vocabulary Size - The larger the vocabulary, the more trouble the machine has in determining what was spoken and the more trouble the user has using and remembering the vocabulary items. The problem is that it would be difficult to justify the expense associated with implementing ASR in most applications if the vocabulary was limited to an unreasonably small size. The effort expended in investigating this variable should be directed toward determining the functional relationship between the size of a vocabulary and:

1. The machine recognition performance.
2. The user's ability to efficiently learn the vocabulary.
3. The user's ability to effectively use the vocabulary once he has learned it.

## CANDIDATE SYSTEMS FOR ASR IMPLEMENTATION

The vocabulary items that were chosen for the experiment were derived to meet the constraints of the experimental design. These items were drawn from a list of items that represent critical F-4 cockpit functions. Each F-4 system and system function was considered and rated as to the appropriateness of ASR for system activation. The systems were rated by four panelists: three pilots and one radar intercept officer. The rationale for assigning priorities centered upon activities which tax the pilot's manipulative ability within the context of the F-4 cockpit configurations and/or his cognitive capacity. That is, activities which tax his capacity for focusing his attention on a single task by requiring that he process information from more than one source at the same time. Many of the functions selected are critical during emergency situations, and it is those situations where voice activation would be particularly advantageous. Another area of consideration is tasks that are time and attention consuming but not necessarily emergency in nature (e.g., radio settings). In situations where the pilot is primarily concerned with the control of the aircraft, primarily during landings and takeoffs, there are tasks which must be performed which are actually secondary to the control tasks. Voice implementation of these secondary tasks would allow the pilot to remain in visual and tactical contact with the control environment while performing the secondary tasks in a quick and efficient manner. Listed below are the results of the systems analysis effort, including a brief synopsis describing why each system was considered appropriate:

a. Emergency release of external stores

Function: Clear all external stores from the aircraft. Used only in the most dire of emergencies such as an engine flameout off the catapult.

Priority Rationale: This function would allow the pilot to get rid of all the external stores while attempting to maintain control of his aircraft.

b. Jettison the centerline station

Function: Jettison the centerline station, usually a large fuel tank. Typically used only in emergencies or imperative combat situations.

Priority Rationale: The primary use of a verbal command function of this sort would be when there is an emergency involving problems with the centerline tank. For instance, upon catapult release the bridle slaps the underside of the centerline tank and a fire erupts. The pilot must immediately jettison the tank and still manage to keep his craft airborne.

c. Mission control: System selection

Function: Select the radar lock up mode for a particular encounter.

Priority Rationale: These commands are used in "the heat of battle" and opportunities for their use are fleeting. Verbal selection would hasten radar locks, potentially resulting in more successful missions.

d. Deploy the drag chute

Function: Deploy the drag chute to slow the aircraft after landing or to help the pilot regain control of the aircraft in certain inflight situations.

Priority Rationale: When the situation is right for deploying the drag chute, the pilot also finds himself in a situation that requires him to devote total cognitive and psychometer capacities to control the aircraft. Deploying the drag chute is an annoying, time consuming and potentially dangerous activity at that particular time.

e. Release the hook

Function: Move the hook into a position to catch an arresting wire upon touchdown.

Priority Rationale: A potentially useful voice call in an emergency or time critical situation (i.e., the pilot forgot to lower the hook and is now making his final approach when the LSO informs him that the hook is in the up position). In addition, under normal operations the hook down type call is usually voiced as part of a checklist or in response to a tower request.

f. Selective missile jettison

Function: Jettison a missile at a particular station due to missile malfunction.

Priority Rationale: The best scenario to argue for voice control is a combat situation where a missile has been activated but malfunctions. A missile that hang-fires or misfires is a dangerous item and must be dealt with immediately. Because speed is important, a verbal command seems appropriate to this situation.

g. Set gunsight

Function: Adjust the positioning of the optical gunsight according to the type of mission being conducted.

Priority Rationale: The benefit of making this function a voice command would occur in transitioning from bombing settings to air-to-air settings. During the Viet Nam conflict, MIG kills were lost due to improper sight settings.

h. Lower (Raise) the landing gear

Function: Raise or lower the landing gear.

Priority Rationale: During any emergency, when the aircraft is either departing or landing, getting the gear up or down in an expeditious manner can mean the difference between a successful mission and a disaster. During landing, a checklist is always read aloud in the cockpit or "gear down" is called to the tower. This action could be used to assure that the gear are indeed down. On takeoff, a verbal "gear up" command would relieve the pilot from having to release his control to activate the gear handle.

i. Probe control: Refuel all tanks

Function: Extend the refueling probe during an air tanking operation where the fuel situation is not immediately critical.

Priority Rationale: The pilot's attention is required for controlling the aircraft. Finding the probe switch requires the pilot to divert this attention from aircraft control.

j. Probe control: Refuel feed tanks first

Function: Extend the refueling probe during an air tanking operation where the fuel situation is critical and fuel is immediately required in the feed tanks.

Priority Rationale: Similar to the refuel all tanks probe control except the situation is more critical. The aircraft is extremely low on fuel and tanking must be successful very soon.

k. RAT control

Function: Extends the RAM Air Turbine generator for emergency electrical power. Voice activation would be useful only if the pilot has some advance warning of impending electrical failure.

Priority Rationale: During an inflight emergency, it may be desirable to provide a backup source of electrical power without having to divert the pilot's activity from that of efficiently coping with the emergency.

l. Jettison the external tanks

Function: Jettison the external fuel tanks from the aircraft to gain a rapid reduction of weight or drag.

Priority Rationale: Of primary importance during emergencies when it is important to reduce weight or drag (e.g., single engine failure during take-offs). The command also would be quite useful for coping with an unexpected combat engagement.

m. Radio: Frequency (Channel) control

Function: Change the frequency (channel) of the aircraft's radio.

Priority Rationale: Multiple frequency (channel) changes can occur during departure and approach (i.e., close to the ground). It is both distracting and vertigo inducing for the pilot to have to look away from his control environment to change the frequency (channel) of the radio.

n. Radio: Automatic Direction Finder (ADF): Frequency (Channel)

Function: Control the ADF frequency (channel) for navigational purposes. Indicates relative bearing of and homes on radio signal sources.

Priority Rationale: Some signals (e.g., emergency beepers) may be received weakly, intermittently or momentarily. A voice command would configure the radio for quick ADF operation, without demanding the distraction of manually setting up to find the signal or the usual delay factor assocaited with manual setup.

o. Change TACAN channel

Function: The Tactical Air Navigation System indicates bearing and distance to ground station, determines identity and dependability of beacon and determines distance to other aircraft.

Priority Rationale: Primarily a convenience function, but it could be important as a voice function in situations such as cross-checking unreliable TACANS or for approaches/departures which require references to more than one station, at a time, when the pilot is busy flying the profile.

p. Radar control

Function: There are a number of controls the NFO must operate to effectively utilize the radar gear. Some of the functions that may be selected include range, pulse doppler, pulse, high map, low map, coverage, and groundspeed.

Priority Rationale: Radar Scopes are said to be hypnotic, especially under pressure (e.g., a "live" intercept run). Verbal selection of radar features would reduce the necessity to look away from the scope and thus diminish the distractions and the need to "get your bearings" on the scope again.

q. Select IFF code (Identification Function)

Function: Select the operation frequency (or IDENT) which the IFF transponder will broadcast.

Priority Rationale: IFF changes, like radio frequency changes, occur frequently on approach and departure, when the pilot is busiest.

r. IFF: System selection: Secure

Function: Secure the transmission of IFF signals.

Priority Rationale: Voice function useful in securing the system prior to combat, a period when the pilot's workload is typically high.

s. Couple the Data Link

Function: Provides data link control to the ground station.

Priority Rationale: The coupling action usually takes place when the aircraft is close to the ground. As a manual function the pilot must focus his attention on the Data Link switch and manipulate the switch. A voicing of this function would allow the pilot to concentrate his visual and tactical resources on controlling the aircraft.

t. Altimeter Setting

Function: Reset the aircraft's altimeter to the current level of the airfield. This is a calibration function to assure accurate reading of the aircraft altitude.

Priority Rationale: A voice command would assure that the altimeter has been reset. When the tower gives altimeter information, the pilot must repeat it back. A voice operated reset function could key on the pilot's voicing and automatically reset the altimeter.

u. Release flares or chaff

Function: Release flares or chaff in an effort to confuse threat radars or heat seeking weapons. Used exclusively during combat.

Priority Rationale: The advantage of making this a voice call should be an increase in NFO performance level. Currently the NFO must redirect his gaze from the radar display to the flare or chaff switch in order to activate this function. In doing so, the NFO momentarily loses contact with the engagement situation or the threat status. Voice command would eliminate this problem.

# IV. EXPERIMENTAL METHODOLOGY AND RESULTS

Speech is defined as a system of symbology in which the meaning of each symbol is determined by social convention. Given the arbitrary nature of the meaning associated with these speech symbols, an interesting manipulation becomes viable and quite appropriate to our context. There are two forms of the learning-utilization interaction that will take place with the institution of AST in a cockpit environment. First, the experienced aircrew member must learn the new vocabulary, associate each item with a stimulus with which he is already intimately familiar and finally utilize the stimulus-response unit in an appropriate fashion. The second learning-utilization situation involves the student pilot coming into a new environment, where he must not only learn the vocabulary and associate it with the environment stimuli, he must also learn to identify the stimuli before he can utilize the unit in an appropriate way.

The situation referred to above is approximated by manipulating stimulus familiarity in an experimental manner that is analogous to the learning-utilization situation. The arbitrary nature of assigning meaning to speech symbols also allows the experimenter the opportunity to manipulate response meaningfulness and familiarity as variables. Meaningfulness is a variable that may remain constant while the response form is varied. For example, the function that the pilot may wish to accomplish is to lower the landing gear. The phrase that he uses to arrive at that configuration may be "lower the landing gear." It may be that the pilot can voice a phrase such as "gear down" much easier or more consistently, and it is a better choice for the vocabulary. The meaning of each phrase is the same, but the response form and, in some sense, the response familiarity changes. In addition, the meaningfulness of the response can be varied while maintaining the familiarity of the phrase. This sort of manipulation also enables the experimenter to investigate the interaction of learning on the utilization of vocabularies.

## EXPERIMENT 1

Experiment 1 was the first attempt to determine the significant factors that must contribute to the specification of an ASR vocabulary. The focus here was upon user performance in a simulated environment as a function of the variables that we manipulate. In this experiment we were using a paired-associate learning and utilization task to simulate the processes of the pilot in learning and utilizing a vocabulary to control various aircraft functions. Since the environment and the task we presented the subject was by necessity sterile, simplified and artificial, the magnitude of the effects we consider significant may seem trivial to the reader. However, given these conditions and the fact that the pilot may have to engage in extraneous cognitive processes while performing the verbal control tasks, we feel that the rigorous examination of the subject's response latencies at the millisecond level was not only justified but required.

In order to assure that we were getting an accurate assessment of the subject's abilities, the experiment was arranged in a way that did not lead the subject to anticipate a particular stimulus or response. If the subject was allowed to anticipate the correct response, by sequential arrangement, probability matching, or in any other way, the experimenter would not get a meaningful assessment of the subject's cognitive capabilities. Due to a number of limitations, cueing/priming, the traditional method for controlling expectancies was not used in this experiment. Rather, a neutral cue (LaBerge, Petersen and Norden, 1978, and Posner and Synder, 1976), which provided no discriminative information about the display, was used.

Method

Verbal Responses. The verbal responses used in this experiment are presented in Table 1. These responses were selected because of their rankings by a random sample of twenty Logicon employees. There were two dimensions upon which the phrases were scaled, familiarity and meaningfulness. A five point scale was used and the dimensions were dichotimized into ratings of high and low. Phrases which on the average were rated at two or less were assigned to the low category; phrases rated at four or higher were assigned to the high category. The initial list of phrases contained 66 phrases. Two different orderings of the phrases were developed and administered as part of the scaling effort.

Table 1. Experiment 1 - Verbal Responses

| | | Phrase Meaningfulness | |
|---|---|---|---|
| | | High | Low |
| Phrase Familiarity | High | 1. Landing Gear Up<br>2. Drop Landing Gear<br>3. Pop the Chute<br>4. Wheels Down | 1. Hook<br>2. Dump<br>3. Ident<br>4. Hook Down |
| | Low | 1. Stow the Ram Air Turbine<br>2. Refuel Feed Tanks<br>3. Set Gunsight 492<br>4. Emergency Release External Stores | 1. Recover the RAT<br>2. Centerline Jettison<br>3. Extend RAT<br>4. Gunsight 492 |

Stimuli. The stimuli for this experiment included eight words and eight nonwords. The nonwords were actually anagrams of the words since they contained the same letters as the word, with the order rearranged. These four and five letter strings, both word and nonword, are shown in Table 2. In addition to the items shown in the table a set of foils for each item was also developed. The foil set consisted of strings which varied in similarity to the actual item. That is, if the item was salt, a highly similar item would be the anagram slat which shares all the letters with the item. Foils were arranged to contain zero, one, two, three, four and five (for five letter strings) letters in common with the stimulus item.

Table 2. Experiment 1 - Stimuli

| Familiary Stimuli (Words) | Unfamiliar Stimuli (Nonwords) |
|---|---|
| LAMP | MPLA |
| READ | ADRE |
| SALT | LTSA |
| LIVE | VLEI |
| ROCKS | CKOSR |
| STARE | ATRSE |
| PEACH | APHCE |
| TRAIL | RLIAT |

Apparatus. The visual stimuli were presented on an ADM-3 CRT terminal with a 30.48 cm diagonal screen. The CRT was driven by special purpose hardware that interfaced with the Nova 3/12 minicomputer. The computer controlled the presentation of the stimuli and recorded the subject's responses via a button board and a Shure dynamic microphone headset in conjunction with a Threshold Technology VIP-100 Speech Recognition Preprocessor. The stimuli were alphanumerics that were conscribed within a 5 x 7 dot maxtrix that measures 2.1 x 4.0 mm on the face of the screen. The display contained a single four or five letter string which subtended a visual angle 1.5 degrees (1.2 degrees for a four letter display). The screen was mounted at eye level, 45.7 cm from the edge of the table where the subject was seated. The response buttons were 25 mm in diameter and required a 1.0 mm downward displacement to close a microswitch. The buttons were mounted on an inclined plane and positioned on the table in front of the subject.

Subjects. Eight experimentally naive employees of Logicon, Inc. served as subjects.

Procedure. A trial began with the presentation of a cue, a string of three upper case X's, in the upper portion of the screen for 500 msec. The cue served to warn the subject that a display was about to be presented. The cue provided no information about the display other than the general temporal information concerning when the display would appear. After a 500 msec blank period the display was presented in the lower portion of the screen for 3000 msec or until a response was made. Responses were recorded by the computer during the display period. If an inappropriate response was made to the display, "WRONG" was presented in the center of the screen for 500 msec. The time between a display and the next cue, the intertrial interval, was 500 msec.

The subject's primary task in this experiment was to voice the proper phrase when a particular target display was presented. Occasionally the item presented as the display was an item for which the subject had not learned to associate a phrase. When this foil display occurred, the subject's task was to press one of the two response buttons. When the subject made a proper response, either by voicing the correct phrase or depressing a response button, the reaction time of the response was presented in the center portion of the screen.

Each experimental session consisted of eight blocks of trials: a voice training block, three learning blocks and four test blocks. The first block was always a voice training block where the subject voiced the phrases and the speech recognition system recorded a reference pattern for the phrase. The voice training block was very similar to the test block described above with the following exceptions:

a. The neutral cue of the test block was replaced with a cue that told the subject which display he would see next and what phrase to voice. The sequence was, for example:

Cue: When you see <u>READ</u> say <u>WHEELS DOWN</u>.
Display: READ

b. There were no foils during this block nor was there any feedback to the subject, other than the removal of the display when the machine had accepted his voicing as a referent.

The second block of the session was a learning block. The learning block was identical to the voice training block except that now, instead of building voice reference patterns, the speech recognition system was used to evaluate the subject's voicings and provide feedback concerning the accuracy and speed of the response. The third block was a test block as originally described. The learning blocks and test blocks were then alternated throughout the session. Each subject participated in the experiment for seven days with one experimental session being conducted each day.

Each test block contained 80 trials: 64 target displays and 16 foil displays. The learning and voice training blocks contained only the target display trials. One half of the target displays were words and the other half were nonwords. The foil items were varied from test block to test block. Each of the sixteen target displays was repeated four times within a block of trials while each of the foil displays appeared only once. The sequencing of the test blocks was balanced across subjects (with a latin square procedure). The voice training and learning blocks were arranged in a similar manner; however, the foil items were not included in either of these blocks. All trials in each of the three types of blocks were presented in a random order.

## Results and Discussion

In this experiment we systematically manipulated three variables: phrase familiarity, phrase meaningfulness and stimulus familiarity. The results of these manipulations were examined in terms of the subject's performance in learning and utilizing the vocabulary items. The learning aspects of the task occurred during the first three days of the experiment. Completion of the learning phase occurred when the adjusted percent correct (accuracy) data approached asymptote.

The adjusted percent correct data were derived from the actual (composite) percent correct data. The actual data represent two sources of error: error that can be directly attributed to the subject and error that can be attributed to the recognition device. Our major concern here was the effects of our experimental manipulations upon responsiveness of the human subject. If the subject voiced the proper phrase but was not recognized by the machine, that should not be held against the subject's performance. In this experiment we did not have the resources to have a person monitor the subject's responses and compare each response to the machine's assessment of the response. We do, however, have a rather direct way of assessing the machine's ability to recognize the subject. In our learning block the cue informed the subject about the stimulus to be presented as the display and about the correct response phrase. Five hundred msecs later the display was presented and the subject responded. In adjusting the percent correct data for machine recognition errors we assume that the percent correct data from the learning blocks represent only the machine error component and not a human error component. Therefore, we argue that the adjusted percent correct score (percent correct test blocks minus percent correct learning blocks) represents a better estimate of "true" human ability.

The utilization of the vocabulary items occur as an experimental phase after the learning phase is complete, from day three through day seven. The utilization of the vocabulary items is measured by the efficiency with which the subject is able to voice his responses. Reaction time (latency) is the primary measure of vocabulary utilization since it is assumed the subjects have reached a high level of accuracy during the learning phase.

The data for both phases of the experiment are presented in Figure 1. In this figure the latency and adjusted accuracy data are presented for high and low familiar phrases, high and low meaningful phrases and familiar and unfamiliar stimuli across days. An analysis of variance of the adjusted percent correct data for the learning phase revealed a significant main effect for: practice across days, $F(2,14) = 32.66$, $p < .01$ and for stimulus familiarity, $F(1,7) = 5.63$, $p < .05$. None of the interactions reached statistical significance; however, there were two that displayed interesting trends toward significance. The interaction of days and stimulus familiarity tended toward an effect, $F(2,14) = 2.86$, $p < .10$, as did the interaction of days (practice), phrase familiarity and phrase meaningfulness, $F(2,14) = 3.67$, $p < .10$. The first interaction trend, days and

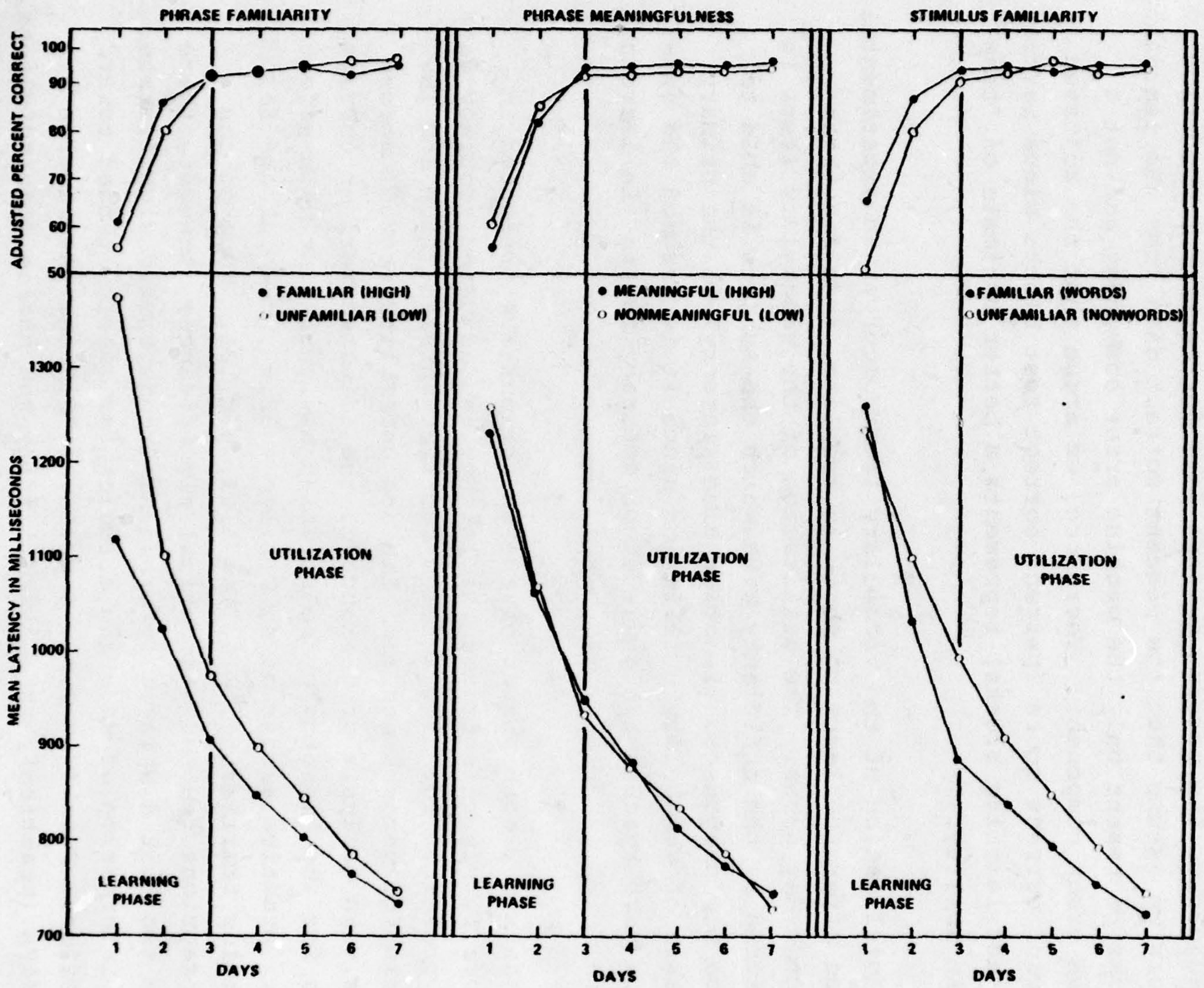


Figure 1. Data from the Vocabulary and Learning Phase of Experiment 1

stimulus familiarity, appears to suggest that the overall accuracy difference for these items is gradually decreasing, that is, the unfamiliar stimuli (nonwords) initially elicit much poorer performance than the familiar stimuli (words). As the training continued, this difference decreases until, as we shall see in the utilization phase, the difference disappears. The second interaction, days, phrase familiarity and phrase meaningfulness, appears to be due to the initially superior performance of high familiar, low meaningful phrases. This superior performance gradually decays relative to the other levels until by day three it is the poorest of the four levels. It should be noted that the phrases that comprise the high familiar, low meaningful were the shortest phrases used in the experiment, perhaps indicating that phrase length may have contributed to this trend.

An analysis of variance of the latency data for the learning phase revealed the following significant main effects: days (practice), $F(2,14) = 33.41$, $p < .01$; and phrase familiarity, $F(1,7) = 7.89$, $p < .05$. In addition, the interaction of days and phrase familiarity was also significant, $F(2,14) = 3.86$, $p < .05$. These effects seem to be focused around phrase familiarity, indicating that the effect is changing as a function of practice. Visual inspection of Figure 1 reveals that on day one the familiar phrases were voiced much quicker than unfamiliar phrases, but that the difference was significantly smaller by day three (response learning). The unfamiliar phrases seemed to become functionally familiar, just as one would expect.

As for the utilization phase data, an analysis of variance of the latency data revealed a significant main effect for days (practice), $F(4,28) = 10.42$, $p < .01$, and a significant interaction effect for days and stimulus familiarity, $F(4,28) = 2.94$, $p < .05$. This interaction indicates that the subjects were learning to treat the nonwords more like the familiar words the more frequently they encounter them. This would seem to indicate that the difference between word and nonword items is just a matter of perceptual learning. The adjusted percent correct data for the utilization phase showed only one significant effect, the interaction of phrase familiarity and phrase meaningfulness, $F(1,7) = 5.81$, $p < .05$. Inspection of this interaction reveals that the high familiarity, low meaningfulness phrases were responded to much less accurately than the other phrases. Again, it must be pointed out that these phrases were shorter in length than the other phrases; therefore, it seems that phrase length should be investigated to determine if this complicated interaction may be due to a differential phrase length effort.

In general, it appears that phrase meaningfulness did not affect the learning or utilization of the verbal material in a direct way. Phrase familiarity, on the other hand, directly affected the learning of the paired-associate items, while only marginally influencing the utilization phase. Stimulus familiarity, on the other hand, directly affected both the learning and utilization phrases. These results would appear to support the need for pretraining of both the stimulus and the response for optimal learning and utilization of a vocabulary item. This conclusion may be a bit overdrawn on the stimulus side, given that the familiar stimuli were words with a host of images and associations connected to each one. One way to eliminate this imagery interpretation is to pretrain a set of nonwords in a way that the subjects become perceptually familiar with them. In the next experiment this sort of pretraining of the stimuli was undertaken and the effect of phrase length was examined in more detail.

## EXPERIMENT 2

The results of Experiment 1 led us to two interesting questions. The first question involves the length of the voiced phrase; more explicitly, are short phrases more difficult for subjects to utilize accurately than long phrases? The second question concerns the stimulus familiarity effect demonstrated in Experiment 1. In Experiment 1, words were used as the familiar stimuli, and nonwords were used as unfamiliar stimuli. Phrases associated with word stimuli were learned faster and utilized more efficiently than phrases associated with the nonwords. This pattern of results may have been due to the perceptual familiarity of the words or to the rich semantic network and imagery associated with each word. For our purposes it seems that the semantic aspects of the stimuli are not an interesting focus, since we are attempting to simulate the perceptual conditions of a senior pilot recognizing a particular cockpit situation and making a voice command as a response to that situation. We will have more to say about this in a later discussion. Perhaps a more direct way to simulate this situation is to perceptually train a group of unfamiliar nonword items until the subjects respond to these stimuli in the same way that they regard word items in a task that is primarily perceptual. A simultaneous matching task has been proposed by some researchers as a primarily perceptual task (LaBerge, 1973; LaBerge, Samuels and Petersen, 1974; Murmurcek, 1977; Petersen and LaBerge, 1977 and Posner and Snyder, 1977). The perceptual learning of the nonword items was accomplished in a pretraining task where subjects were asked to make same-different judgments of simultaneously presented words and nonwords. This training task was conducted over five days with three sets of stimulus items: words, nonwords that were repeated each day, and nonwords that were new each day. At the beginning of the experiment, it is expected that the word items will be responded to faster than both types of nonword items. However, as the exposures to the repeated nonword items increase, the reaction time to these items should decrease relative to the other two items until by the last day of training these repeated nonwords are responded to at the same rate as the words, while the novel nonwords remain at about the same relative level as they were on day one. The convergence of the latencies of the repeated nonwords, with that of the words, can be taken as evidence of perceptual learning. In this manner we hoped to develop stimuli that were perceptually familiar yet did not have a large array of semantic associates and imagability built in. The next step was to test these stimuli in the paired-associate task used in Experiment 1.

## Perceptual Learning Pretraining Methodology

Stimuli. The stimuli for this experiment included two four-item sets of words and two four-item sets of nonwords that were repeated from day-to-day, and five four-item sets of nonwords that were not repeated from day-to-day. The repeated nonwords and the novel nonwords that were used on day one and day five were anagrams of the word items. The stimuli are presented in Table 3.

Apparatus. The same apparatus used in Experiment 1 was used in this experiment. The display consisted of two four-or-five-letter strings which were separated by 101 cm, corresponding to 40 spaces on the screen. This type of display subtended a visual angle of 12.46 degrees.

Subjects. Eight experimentally naive employees of Logicon, Inc. volunteered to serve as subjects.

Table 3. Experiment 2 - Stimuli

| Words | | Repeated Nonwords | |
|---|---|---|---|
| Set A | Set B | Set A | Set B |
| LAMP | SALT | MPLA | LTSA |
| READ | LIVE | ADRE | VLEI |
| ROCKS | TRIAL | CKOSR | RLIAT |
| STARE | PEACH | ATRSE | APHCE |

| | Nonrepeated Nonwords | | | | |
|---|---|---|---|---|---|
| Day: | 1 | 2 | 3 | 4 | 5 |
| Set A: | AMPL | VNEO | DTAE | KLMI | MLPH |
| | AERD | LPEH | NESD | NDEI | AEDR |
| | KSROC | GLHTI | SLRU | CHBNE | RSKOC |
| | ETSRA | DBRIA | DAEP | ESNES | TSAER |
| Set B: | ASLT | KLMI | VNEO | PTAE | TLSA |
| | IEVL | NDEI | LPEH | NESD | ELLV |
| | ATLRI | CHBNE | GLHTI | ESLRU | ILRTA |
| | CPAEH | ESNES | DBRIA | RPAEP | AEPHC |

Procedure. The procedure used in the perceptual learning pretraining phase of Experiment 2 was quite similar to the procedure used in Experiment 1 with the following exceptions. First, the duration of the display period was 1500 msec. Second, the subject's task was to look at the display and determine if the items displayed matched one another. If they did match, he was instructed to depress the right button, and if they did not match, he depressed the left button.

Each experimental session (one day each) contained four blocks of trials, each involving the same simultaneous matching task and the same target display trials. Foil display trials varied from block to block and were balanced across subjects (using a latin square method). Each block contained 60 trials: 48 target displays and 12 foil displays. Sixteen of the trials were word displays (each word was repeated four times) sixteen were repeated nonwords and sixteen were novel nonwords. The same ratio applied to the foil displays. All trials were presented in random order within a block.

## Results and Discussion

The data presented in Figure 2 are the results of the perceptual learning pretraining portion of Experiment 2. These data depict the relative changes in stimulus familiarity as a function of days (practice) for three levels of stimuli: words, repeated nonwords and novel (nonrepeated) nonwords, in terms of the subject's response accuracy and latency.

For purposes of a clear description of the results, the relative changes in mean response times are plotted for the latency of novel nonwords minus the latency for words, and the latency of repeated nonwords minus the latency for words as a function of days. An analysis of variance of the latency data for this pretraining phase of the experiment reveals significant main effects for days (practice), $F(4,28) = 17.55$, $p < .01$, and familiarity, $F(2,14) = 50.92$, $p < .01$. In addition, the days by familiarity interaction also was significant, $F(8,56) = 2.88$, $p < .01$. An analysis of variance of the accuracy data found no significant main effects or interactions. The essence of this data can be seen most easily by inspecting the difference curves. These curves represent the differences in reaction time of the two nonword conditions and the word condition. The assumption is that the word condition represents the most efficient level of processing possible, since the subjects are intimately familiar with these particular strings of letters. The nonwords, on the other hand, are new to the subjects, so the difference in reaction times for the nonwords relative to the words represents the degree of unfamiliarity of the nonword items. It is notable in

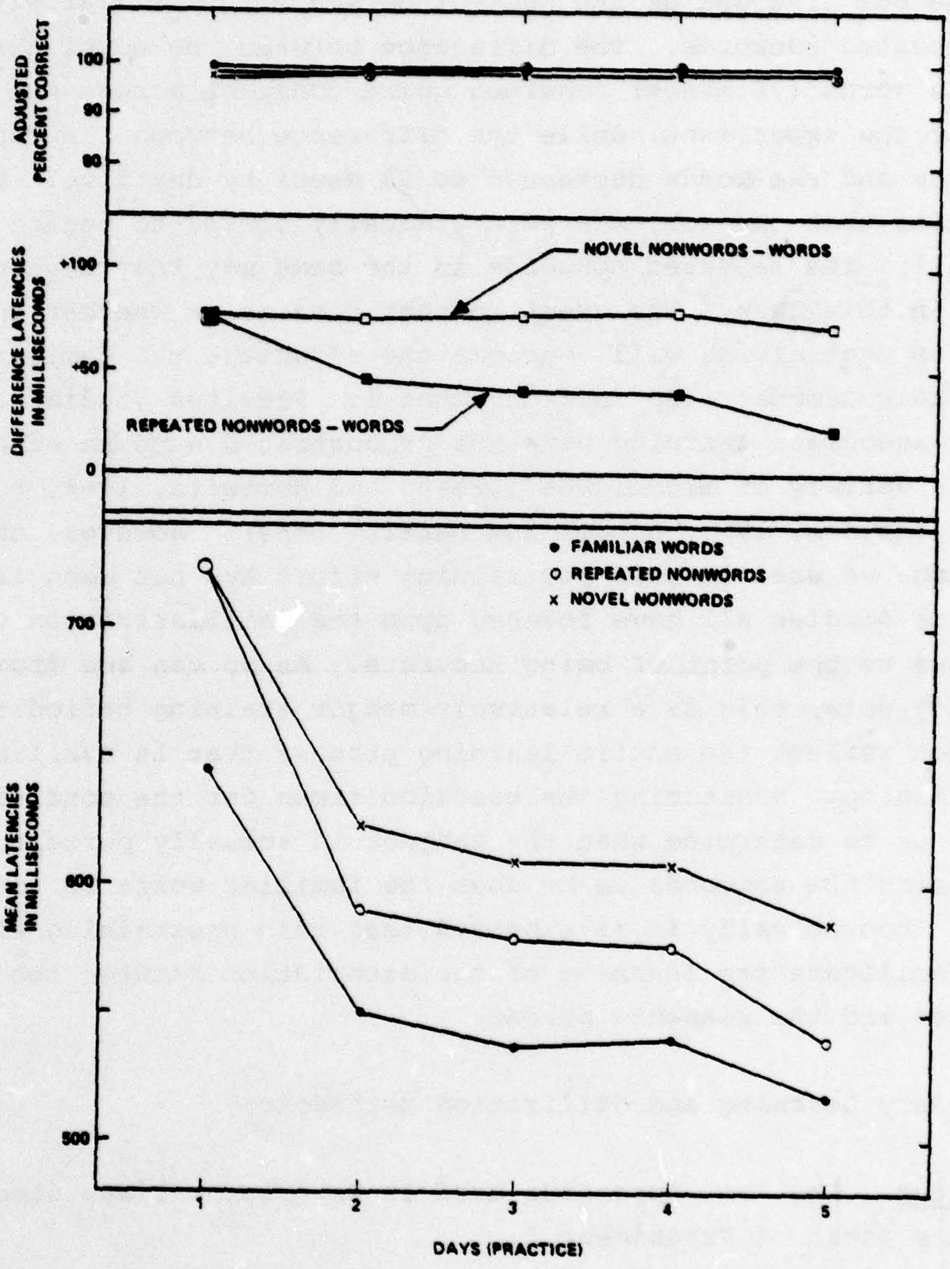


Figure 2. Data from the Perceptual Learning Pretraining for Experiment 2

the difference curves that both nonword sets started together on day one but diverged as the subject became more familiar with the repeated nonwords. The difference between the novel nonwords and the words (76 msecs) remained quite constant across the five days of the experiment, while the difference between the repeated nonwords and the words decreased to 23 msecs by day five. This indicates that the subjects were gradually coming to regard (perceptually) the repeated nonwords in the same way they regarded words in this task. The question that remains is whether this stimulus pretraining will overcome the advantage the familiar word items demonstrated in Experiment 1. Previous studies of paired-associate learning have not demonstrated such an effect using a variety of techniques (Greeno and Horowitz, 1968, Postman and Greenbloom, 1967, Schulz and Martin, 1964). However, the technique we used in this pretraining effort has not been tried. Previous studies all have focused upon the familiarization of the stimulus to the point of being accurate. As we can see from the accuracy data, this is a relatively meager training period that does not reflect the entire learning process that is available for training. Monitoring the reaction times for the conditions allows us to determine when the subject is actually perceptually processing the nonwords as he does the familiar words in this task. Theoretically it is expected that this pretraining effort will facilitate the learning of the association between the stimulus and the response phrase.

Vocabulary Learning and Utilization Methodology

Apparatus. The same apparatus used in Experiment 1 was used during this phase of Experiment 2.

Subjects. The eight subjects who participated in the pretraining portion of the experiment continued their participation in this phase.

Procedures. The same procedures used in Experiment 1 were in effect here with the following exceptions. Each test block contained 60 trials: 48 target vocabulary trials and 12 foil button response trials. One-third, 16, of the target trials contained word stimulus, one-third contained repeated nonwords, and the final third contained novel nonwords.

Verbal Responses. Two sets of verbal responses were generated for this experiment and are presented in Table 4. Each set contained six short phrases and six long phrases. Short phrases contained three distinct syllables or less, while long phrases contained five syllables or more.

Table 4. Experiment 2 - Response Phrases

| | | |
|---|---|---|
| Set A: | Stop Dump | Identification System Off |
| | RAT Out | Station Two Jettison |
| | Gear Up | Emergency Jettison |
| | Squawk Ident | Stow the Emergency Generator |
| | Dump Wing Fuel | Centerline Jettison |
| | Hook | Gunsight 847 |
| Set B: | Wheels Down | Release External Tanks |
| | Stop Fuel Dump | Jettison Station Two |
| | Refuel All | Set Gunsight 847 Mils |
| | Drag Chute | Centerline Stores Jettison |
| | Stow RAT | Extend the Ram Air Turbine |
| | Dump | Emergency Release External Stores |

Results. The second phase of Experiment 2 was analyzed in a manner congruent with the analysis of Experiment 1. Two variables were considered in the analysis, phrase length and stimulus familiarity. The data from both the learning and utilization phases of Experiment 2 are presented in Figure 3. In this figure the latency and adjusted accuracy data for each level of both variables are plotted across days. An analysis of variance of the adjusted accuracy data for the learning phase found two main effects to be significant: days (practice), F(2,14) = 50.57,

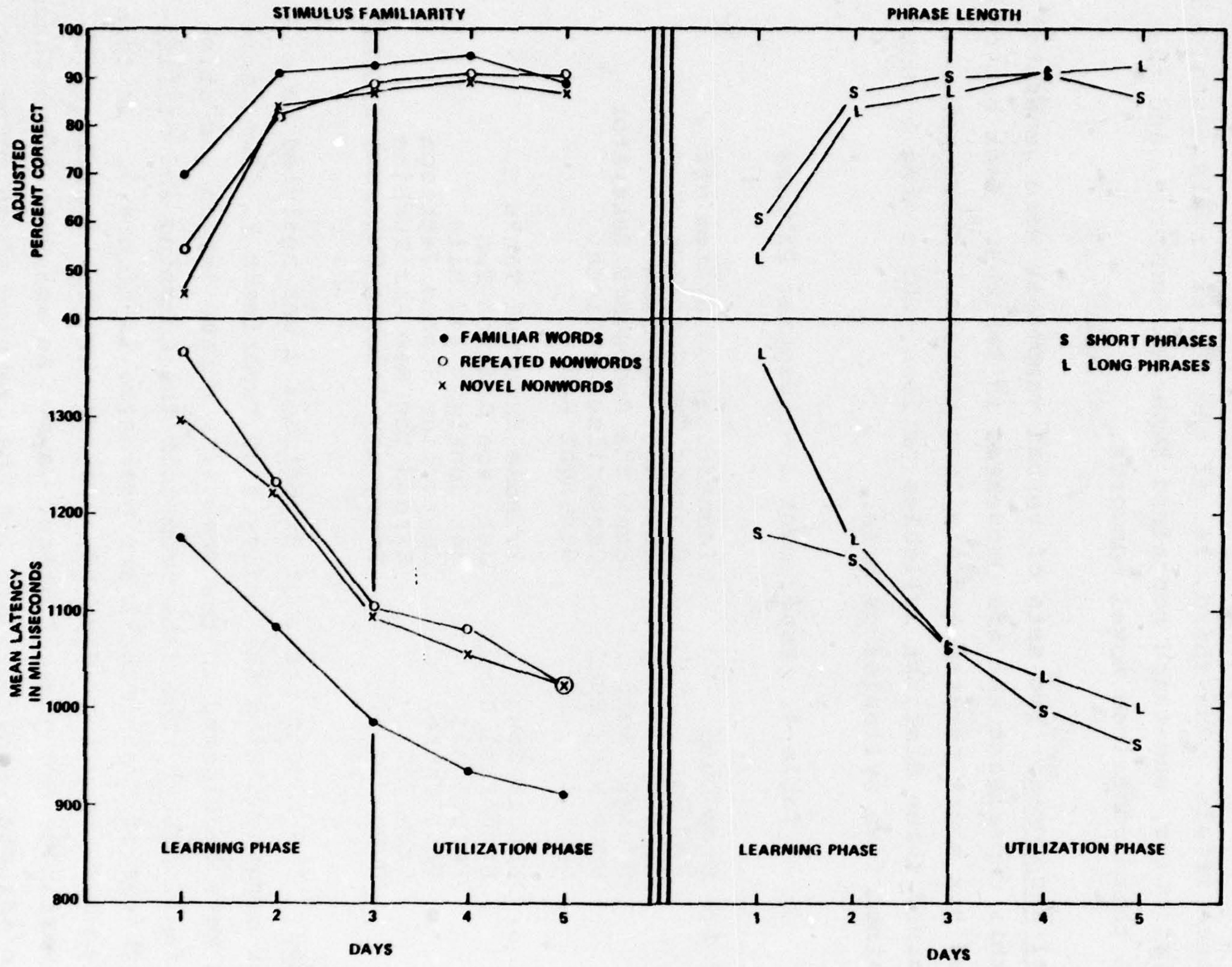


Figure 3. Data from Experiment 2

$p < .01$; and stimulus familiarity, $F(2,14) = 10.37$, $p < .01$. The interaction of days and stimulus familiarity was also found to be significant: $F(4,28) = 2.97$, $p < .05$. The stimulus familiarity effect indicates that in terms of an overall difference, the perceived difference between words and the nonword conditions is real; words are responded to more accurately than either of the nonwords. This suggests that the stimulus familiarity effect observed in Experiment 1 may have been due to more than just the perceptual familiarity of the stimulus items (words). Furthermore, the significant interaction of practice and stimulus familiarity indicates that the advantage of the words relative to the nonword conditions, and the advantage of the repeated nonwords over the novel nonwords, are both quite fleeting. An analysis of variance of the latency data from the learning phase also revealed two significant effects: days (practice), $F(2,14) = 10.68$, $p < .01$; and stimulus familiarity, $F(2,14) = 3.88$, $p < .05$. No significant interactions were encountered. It appears that the word items were responded to more quickly than the nonword items. In addition, there were two large, but apparently artifactual, differences that did not reach significance. On day one the repeated nonwords were apparently slower relative to the novel nonwords, and although the difference did not persist it was entirely contrary to the predicted result of facilitation for the repeated nonwords. The phrase length data also appeared to have exhibited a difference on day one. Long phrases appear to be responded to much slower than short phrases. This result is consistent with previously reported data (Johnson, 1978), but the transiency of the difference has not been reported before. This day one difference of 183 msecs was not significant even by t-test, perhaps suggesting a lack of power in this experiment due to the relatively small sample size.

As for the utilization phase, an analysis of variance of the latency data gain revealed two significant effects and no significant interactions. The main effects were: days (practice), F(2,14) = 7.48, $p < .01$ and stimulus familiarity, F(2,14 = 3.51, $p < .05$. The familiarity effect is obviously due to the superior performance of the word condition relative to the nonword conditions. Interestingly, the convergence of the word and nonword items, so apparent in Experiment 1, was not displayed here. Again, no differences could be demonstrated for the phrase length conditions. The analysis of the adjusted accuracy data for the utilization phase found no significant main effects or interactions.

# V. CONCLUSIONS AND RECOMMENDATIONS

This section provides discussion of three important topics for cockpit ASR application. The first is a theoretical discussion of the results of the experiments, including implications for using ASR in aircraft cockpits. The second topic concerns the problem of defining a vocabulary in a manner that significantly reduces the probability of recognition confusion on the part of the machine or the human speaker. The last topic is our recommendations for further work in applying ASR to cockpits.

## THEORETICAL DISCUSSION OF THE EMPIRICAL FINDINGS AND IMPLICATIONS FOR COCKPIT ASR APPLICATIONS

This study examined several aspects of verbal behavior with the orientation of determining viable considerations for specifying the important characteristics of the vocabulary phrases for ASR applications. Four variables were examined with regard to their influence on the subject's performance in learning and utilizing vocabulary items. The most important variables appeared to be the speaker's level of familiarity with the phrase and his level of familiarity with the environment in which the vocabulary will be used. The other two variables investigated were the meaningfulness of the phrase and the length, in syllables, of the phrase; neither of these variables could be shown to have a consistent effect on either the speaker's learning or utilization of the phrases.

Phrase familiarity exhibited its major influences during the learning phase of the experiment; however, looking at the overall pattern of results for this variable, the results seem quite consistent. Initially, unfamiliar phrases are responded to less accurately and much slower than the familiar phrases. The accuracy difference disappears during the first three days of training, but the latency difference persists. However, this latency difference gets smaller as training progresses, until by the seventh and last day the unfamiliar phrases are used three milliseconds faster than the familiar items. The point is that using phrases which are familiar to the speaker will facilitate his performance for both the learning and utilization of the phrase. However, if the use of unfamiliar phrases is necessary, it seems likely that a pretraining procedure should be beneficial.

This same type of pretraining argument was raised for stimulus familiarity in Experiment 1 and investigated in Experiment 2. The investigation pointed to an area where our knowledge of human cognition begins to wane. The data from Experiment 1 demonstrate a convergence of stimulus familiarity conditions in terms of accuracy during the learning phase and in terms of latency during the utilization phase. These data are very similar to the data for phrase familiarity. The data suggest that the more familiar the subject is with the stimulus items, the easier it will be for him to learn and utilize the vocabulary phrases.

The pretraining argument was made, and the verbal learning literature was reviewed. The literature provided no support for the stimulus pretraining argument. In fact, the literature contains several studies that report that pretraining did not affect the subsequent learning tasks. These studies, however, had only pretrained the stimulus to the level of accurate performance. The data from Experiment 1 suggest that more learning can be accomplished than is reflected by the accuracy data. The latency data continued to show a familiarity effect during the utilization phase. This effect began to gradually disappear, indicating that the subjects were gradually learning something more about (chunking) the unfamiliar stimulus items.

A pretraining phase was attempted in Experiment 2. Stimulus familiarity was systematically trained using strings of letters that represented varying degrees of familiarity to the subjects. Words were used as the familiar items, and two sets of nonwords were used as the unfamiliar items. During the pretraining the subjects were repeatedly exposed to (familiarized with) one set of nonwords while the other set was not introduced until the last pretraining day.

The data from the pretraining indicated that the subjects had begun to regard, that is, respond to, the repeated nonwords in the same way they were responding to the words. From these data it was concluded that the subjects were equally familiar with both the words and the repeated nonwords. Therefore, the stimulus familiarity effect was expected to show up only when the novel nonwords are compared to the repeated nonword for the paired-associate task. Responses to words and repeated nonwords, being equally familiar, would show no discernable difference.

When the subjects were asked to learn to voice particular phrases when the stimulus items appeared (paired-associate learning) a rather curious thing happened. The subjects did not respond to the repeated nonwords in the same way as they did the familiar words. In fact, these pretrained nonwords were responded to much less accurately and much slower than the words. One explanation for this may be that the words are much easier to form images of or use in imaginable mneumonics than are the nonwords, and perhaps this or a similar reason can account for the difference. We should really be looking at the difference between the repeated nonwords and the novel nonwords. If pretraining was beneficial, then repeated nonwords should be responded to more accurately than the novel nonwords. That prediction was borne out on day one, but the difference completely disappeared on day two and beyond. The latency data, however, is where the pretraining argument encountered its major problems. The latencies for responses to the repeated nonwords were much slower than responses to the novel nonwords on day one and the same from there after. Pretraining did *not* seem to benefit the learning or utilization of the verbal responses. While the differences found in the accuracy and latency data for day one may be real, the chances are good that they resulted from a differential speed-accuracy tradeoff. The conclusion that must be drawn is that stimulus pretraining does not facilitate the learning, and it appears that *stimulus familiarity is quite likely to be ineffective in influencing subsequent learning*.

This conclusion is not consistent with most theoretical descriptions of the learning process, particulary the ever popular associationist notion of stimuli being directly associated with responses. It is usually argued that the better known the stimulus is, the easier the association should be to learn. The coding theorists (e.g., Johnson, 1972; Estes, 1972) also have trouble with this conclusion; even though their associations are mediated by a higher level code, they still require some easily recognizable representation in memory of the stimulus. This representation can then be associated to the higher level code in a direct way. The problems these theorists encounter with the data presented above can be overcome by adopting a contextualist orientation to the theories (e.g., Jenkins, 1974; Estes, 1976; LaBerge, 1977; Petersen and LaBerge, 1977). The contextualist orientation provides for multiple codes to be derived for each stimuli. Each of these codes results from the perceptual learning that occurs in a particular context. Hence, the learning that occurred during the preceptual pretraining of

Experiment 2 may have resulted in a perceptual code for the repeated nonwords that was totally inappropriate to the paired-associate task; therefore the repeated nonwords were responded to in the same way as the novel nonwords. Another possibility is that the subjects attempted to use this inappropriate, though highly learned, perceptual code the first few times the stimuli were presented. The result may have been that the repeated nonwords required more time for an appropriate code to be accessed, since there may have been competition from the perceptual code constructed during the pretraining.

If this contextualist view of the learning process is at all adequate, the data for the word stimuli also should be handled by the theory. The contextualist view would assume that since words are encountered so often in every day life, there no doubt have been a large number of codes developed for these items. So, not only was there a handy code available for use in the pretraining task, there was quite possibly another code available that was appropriate to the paired-associate task. Therefore it is quite possible that the word items were familiar in both contexts through different representations.

Where does all that leave us with regard to the question at hand? What is the role of stimulus familiarity in specifying an ASR vocabulary? It seems we should qualify our conclusion to include a contextualist interpretation, at least until we can resoundingly refute it. Familiarity with a stimuli in a given context may have an effect on both the learning and utilization of phrase voicings that are made as responses to that stimuli. Training pilots to use a phrase as a control function may not transfer fully to an actual cockpit unless the stimuli used to elicit the phrase during training are provided in the proper context, i.e., the cockpit context. That is not to say that no benefit will be derived from training in an artificial environment. Phrase learning will occur, and increased phrase familiarity will enhance the learning of the eventual association.

Phrase length was shown in Experiment 2 to be of little influence in affecting the performance of the subjects. There was an interesting trend in this data, however, that may support a comment heard from every aviator who participated in this study. Each of them suggested that short phrases would be much better as vocabulary items. Theoretically it seems to make sense that short phrases would be preferred, since they would be easier to chunk (learn); therefore, the learning of the association should proceed faster. There does not appear to be any reason to believe, theoretically, that associations involving short phrases should be easier to utilize than long phrases, other than the obvious difference that it takes longer to articulate a long phrase than a short one. When we look at the data for the nonword stimuli as a function of phrase length, we find that our theoretical expectations were confirmed, at least for the latency data from the learning phase. Long phrases associated with the repeated (familiar) nonwords were responded to much more slowly (119 milliseconds) than short phrases associated with the repeated nonwords. These data are contrasted with the latency data for long and short phrases associated with the novel (unfamiliar) nonwords, which do not appear to be different (11 milliseconds). The analogy that develops is that responses to the familiar items correspond to the responses of experienced pilots, while responses to the unfamiliar items correspond to the responses of inexperienced pilot trainees. Experienced pilots expressed a preference for short phrases which the trend in the data supports. Although the support from the data is not statistically significant, the difference is quite large and may just require a little more powerful sample (more subjects) to reach significance. Therefore, the designer interested in training experienced pilots may find a vocabulary containing short phrases to be the most satisfactory.

## ACOUSTIC CONFUSABILITY AND VOCABULARY SELECTION

Acoustic confusability refers to the degree of similarity between voicings of two different phrases. Naturally, the manner in which the phrase voicings are encoded directly effects the perceived (realized) similarity between the phrases. Therefore, it is expected that the pattern of confusions made by the recognition machine should be different than the confusions made by a human for a given vocabulary. In order to address this problem quantitatively, we began to develop a methodology that would allow the independent evaluation of these confusions. As the analysis proceeded it became very clear that the solution was not at all trivial or necessarily straightforward. The magnitude of the problem was overwhelming. To deduce a best vocabulary for 12 ASR functions with four alternative candidate phrases each, the number of potential vocabularies that must be considered is on the order of six million. In addition, assuming we could quantify the relative similarity of each phrase with every other phrase, it was not all clear how to choose the "best" vocabulary. In fact, it was not clear that there would be a "best" vocabulary, perhaps just several good vocabularies. These problems are addressed in Appendixes A and B.

These problems have been discussed elsewhere and a very good discussion is found in a Ph.D. dissertation by Robert Goodman (1976). Goodman addressed several topics in his paper including acoustic ambiguity, lexical ambiguity, syntactic restrictions and the combination of vocabulary ambiguity and syntactic complexity. In his discussion of acoustic ambiguity, Goodman limited his concern to the segmenting of the speech signal to arrive at a parametric representation of the acoustic space and the subsequent classification of the segments. Confusability and three methods for relating the conditional probabilities of each segment were discussed in some detail. The three methods discussed were actual counts, acoustic-parametric metrics and theoretical models. Although our concern here is not acoustic ambiguity as Goodman defined it, the three methods for generating conditional probabilities should be considered as alternatives to the methodology presented in Appendix B.

Goodman's second topic, lexical ambiguity, was defined as the "ambiguity that occurs when some word of the vocabulary (lexicon) is confused with another word because the two are phonetically similar." This is precisely what we mean by acoustic confusability in the cockpit application. Goodman's approach to the lexical ambiguity (acoustic confusion) problem was "to find a measure of the complexity of a vocabulary so that two may be compared." His approach was to regard the recognition process as a noisy channel and compute the information loss as a measure of the ambiguity, or complexity, of the system. The mechanics of the approach are treated fairly rigorously and will not be described here, so the interested reader is referred to the source.

This methodology, while very good for determining the complexity or ambiguity of a vocabulary for ASR, may not be adequate when human speech generation and recognition, or automatic speech understanding (ASU), is the process being described. Syntax is an important ingredient for both human cognition and ASU which is not accounted for in Goodman's analysis of lexical ambiguity. The last portion of this section contains our recommendations for future applications research in cockpit ASR implementation. The focus of the recommendations is the role of syntax both from the perspective of the recognition machine and the human speaker.

## RECOMMENDATIONS

In this study we examine some unidimensional aspects of vocabulary selection. While the results of the study were enlightening, they do not provide the comprehensive data needed to rigorously specify an optimal vocabulary, even when coupled with the methodologies discussed above. A tentative vocabulary for ASR control of selected F-4 cockpit functions is presented in Appendix C. In order to specify a less tentative vocabulary, we need a more complete theory of cognition, so that the context and grammer of a vocabulary can be accounted for in defining the contents of the vocabulary. It is the human side of the dialogue where the real uncertainties exist. The machine will improve steadily. In fact, some very important advances have been made recently. A continuous speech recognition system (Nippon Electric Corporation) and a speaker independent system (Dialog Systems, Inc.) have been announced. Although these systems are somewhat limited, they do support the notion that machine recognition will not be the bottleneck in future ASR applications. The problem will be to use the recognition capability in a manner that is maximally effective from the user's point of view. In general, what we need is a complete theory of cognition that includes mechanisms for describing language, symbology, learning, problem solving, belief systems, cognitive development, categorization, etc. (Norman, 1979).

The complete theory is an ideal, of course, and we can and must get along with less. Science proceeds by the process of empirical elimination. In this study we examined a number of variables that held clear promise for our application, but the answers were not complete. Further research is required, but that research needs direction. The direction seems clear; we need a better understanding of how syntax interacts with vocabulary utilization. We need to test alternative forms of syntax with regard to human performance and, to a lesser degree, machine performance. (Although the concern here is automatic speech recognition, automatic speech understanding will be extremely important for future nontrivial automatic speech applications; therefore, the understanding of both artificial intelligence and the human cognition will be essential). As for near term expectations for ASR cockpit applications, parallel efforts must be maintained in the areas of:

a. system adaptation to the changes in speech signals as a function of perturbations from the environment (e.g., excessive G's, the oxygen mask, etc.)
b. natural language processing
c. human factors aspects of the human-computer dialogues
d. integration of commercially practical speech recognition systems into the cockpit environment.

# REFERENCES

Brown, R. and Hildum, D. C. Expectancy and the perception of symbols. *Language*, 1956, 32, 411-419.

Coler, C. R., Plummer, R. P., Huff, E. M. and Hitchcock, M. H. Automatic speech recognition at NASA-Ames Research Center. In R. Breaux, M. Current, and E. M. Huff (eds.), *Proceedings: Voice Technology for Interactive Real-Time Command/Control Systems Application*. NASA, Ames Research Center, Moffett Field, California, 1977.

Conrad, R. Acoustic confusions and memory span for words. *Nature*, 1963, 197, 1029-1030.

Conrad, R. Acoustic confusions in immediate memory. *British Journal of Psychology*, 1964, 55, 75-84.

Conrad, R. Speech and reading. In J. F. Kavanagh and I. G. Mattingly (eds.), *Language by Ear and by Eyes*, Cambridge, Mass.: The MIT Press, 1972.

Curran, M. Voice integrated systems. In R. Breaux, M. Curran and E. M. Huff (eds.), *Proceedings: Voice Technology for Interactive Real-Time Command/Control Systems Application*. NASA-Ames Research Center, Moffett Field, California, 1977.

Estes, W. K. An associative basis for coding and organization in memory. In A. W. Melton and E. Martin (eds.), *Coding Processes in Human Memory*. New York: Wiley, 1972.

Estes, W. K. Memory, perception, and decision in letter identification. In R. L. Solso (ed.), *Information Processing and Cognition: The Layola Symposium*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1975.

Feuge, R. L. and Geer, C. W. *Integrated Applications of Automated Speech Technology*, Final Report. Technical Report, ONR-CR213-158-1AF, 1978.

Goodman, R. G. *Analysis of Languages for Man-Machine Voice Communication*. Technical Report, AFOSR-TR-77-0055, 1976.

Grady, M. W. and Micklin, M. *Use of Computer Speech Understanding in Training: A Demonstration Training System for the Ground Controlled Approach Controller*. Technical Report, NAVTRAEQUIPCEN 74-C-0048-1, 1976.

Grady, M. W., Hicklin, M. B. and Porter, J. E. Practical applications of interactive voice technologies - some accomplishments and prospects. In the Proceedings of Voice Technology for Interactive Real-time Command/Control Systems, Ames Research Center, Dec. 1977.

Greeno, J. G. and Horowitz, L. M. On unitizing a compound stimulus. Journal of Verbal Learning and Verbal Behavior, 1968, 7, 913-917.

Hall, John F. Verbal Learning and Retention. New York: J. B. Lippincott Co., 1971.

Jenkins, J. J. Remember that old theory of memory? Well, forget it! American Psychologist, 1974, Vol. 28, #11, 785-795.

Johnson, N. F. Organization and the concept of a memory code. In A. Melton and E. Martin (eds.), Coding Processes in Human Memory. Washington, D.C.: H. V. Winston & Sons, 1972, pp. 125-160.(a)

Johnson, N. F. Coding processes in memory. In W. K. Estes (ed.) Handbook of Learning and Cognitive Processes, Vol. 6, Hillsdale, N.J.: Lawrence Erlbaum Associates, 1978.

LaBerge, D. Attention and the measurement of perceptual learning. Memory & Cognition, 1973, 1, 268-276.

LaBerge, D. Perceptual learning and attention. In W. K. Estes (ed.), Handbook of Learning and Cognitive Processes, Vol. 4, Hillsdale, N.J.: Lawrence Erlbaum Associates, 1976.

LaBerge, D., Petersen, R. J. and Norden, M. J., Exploring the limits of cueing. In S. Dornic (ed.), Attention and Performance, VI. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

LaBerge, D., Samuels, S. J. and Petersen, R. Perceptual learning of artificial letters. Technical Report No. 6, Minnesota Reading Research Project, University of Minnesota, 1974.

Martin, James, Design of Man-Computer Dialogues. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1973.

Montague, H. Voice Control Systems for Airborne Environments. Technical Report, RADC-TR-77-189, 1977.

Murmurek, H. H. C Processing letters in words at different levels. Memory and Cognition, 1977, 5, 67-72.

Normal, D. Ten issues for cognitive science. Paper presented at the First International Conference on Cognitive Science, La Jolla, California 1979.

Petersen, R. J. and LaBerge, D. Contextual control of letter perception. Memory and Cognition, 1977, Vol. 5(2), 205-213.

Posner, M. I. and Snyder, C. R. R. Facilitation and Inhibition in the Processing of Signals. In P. Rabbitt and S. Dornic (eds.), Attention and Performance V, London: Academic Press, 1975.

Postman, L. and Greenbloom, R. Conditions of cue selection in the acquisition of paired-associate lists. Journal of Verbal Learning ahd Verbal Behavior, 1967, 73, 91-100.

Quartano, A. G. Human Verbal Behavior Considerations in the Design of Voice Actuated Hardware Systems. Unpublished dissertation submitted to the Naval Postgraduate School, Monterey, California, 1977.

Schulz, R. W. and Martin, E. Aural paired-associate learning: stimulus familiarization, response familiarization, and pronounciability. Journal of Verbal Learning and Verbal Behavior, 1974, 3, 139-145.

Tikofsky, L. S. and McInish, J. R. Consonant discrimination by seven year olds: A pilot study. Psychonomic Science, Science, 1968, 10, 61-62.

# APPENDIX A

## ACOUSTIC CONFUSABILITY AND OPTIMAL VOCABULARY SELECTION

This memo is addressed to the problem of choosing an optimal vocabulary for a task environment in which the final vocabulary choice is to be made solely on the basis of interword acoustic confusability. No algorithm description is to be found in this memo. Instead, we restrict ourselves to the process of restricting and clarifying the problem and discussing the difficulties involved in finding a satisfactory solution. A subsequent memo will be devoted to a discussion of one or two algorithmic approaches to the solution of the problem which we describe here.

The general setting with which we are concerned is a task environment in which there are a number of functions to be performed. Each function is to be associated with a single utterance consisting of one or more words. We assume that, for each function to be performed, a number of utterances have already been chosen as candidates for the single utterance to be associated with that function. (For example, if the task environment is an aircraft cockpit, a function to be performed might be the operation of lowering the landing gear, and candidate utterances to be associated with this function might be "wheels down," "lower landing gear," "bring wheels down," "lower wheels," and "drop wheels.") We call the total collection of candidate utterances for all functions the lexicon, and a potential vocabulary is a subset of the lexicon which contains precisely one utterance for each function to be performed.

If, in the task environment, there are N functions to be performed, and if the $i^{th}$ function has $n_i$ candidate utterances associated with it, then the total number N* of possible choices of potential vocabularies is the product of the $n_i$ - i.e.,

$$N^* = \prod_{i=1}^{n} n_i$$

To get a feeling for the magnitude of N*, let's assume that there are 12 functions to be performed and that each function has 4 candidate utterances. Then the total number of potential vocabularies is $4^{12} = 16,777,216$. It soon becomes clear that, except in cases where N and the $n_i$ are all relatively small, there are too many possible potential vocabularies for us to find an optimal one by exhaustive search.

At this point we have still not defined "optimal vocabulary." The number of possible potential vocabularies is independent of whatever measure of goodness of vocabulary that we use. But, of course, we cannot determine the "best vocabulary" until we specify what "best" means. Further, the meaning we should attach to "best" here is not immediately obvious. In a general context, we might choose "best vocabulary" to mean that potential vocabulary whose items are easiest to remember, easiest to pronounce, most meaningful, least confusable to human beings, or least confusable to machine. Ideally, we might like to choose the best vocabulary to embody all the last named features. But choosing an optimal vocabulary based on these several measures of goodness is a rather broad problem whose solution probably requires a very careful consideration of weighting for each of the features.

To simplify this broader problem, we assume that all potential vocabularies are equally acceptable in all respects except for the factor of acoustic confusability between pairs of utterances. Even in this restricted case, where acoustic confusability is the only criterion, the definition we ought to adopt for "optimal vocabulary" is not clear. The reason for this difficulty is that acoustic confusability is, by its nature, measurable only for pairs of words, and a usable definition of "optimal vocabulary" must indicate how to use this pairwise information effectively.

One definition of an optimal vocabulary which suggests itself is that an optimal vocabulary is a potential vocabulary whose maximal interword acoustic confusability is less than or equal to the maximal interword acoustic confusability of all other potential vocabularies.

The problem with this definition is that it distinguishes poorly between potential vocabularies since it gives a measure for the confusability of a whole vocabulary based only on the most confusable pair of items in that vocabulary.

In the memo to follow this one we will offer a definition of optimal vocabulary which seems to be intuitively appealing. It avoids the difficulty involved in the definition above, but it is more technical and is more appropriately described along with the algorithm that will use it.

There is another general issue that should be discussed here. The only data we have available to use to determine an optimal vocabulary are confusability measures of the form $\langle a_n^i, a_m^j \rangle$ where this notation means the machine score of the ith item for the nth function against the jth item for the mth function. Since these numbers are the only basic data we have to use, we must approach them carefully and extract as much information from them as we can. In particular, we should note that there exist some methods of machine scoring for which the confusability measures $\langle a_n^i, a_m^j \rangle$ and $\langle a_m^j, a_n^i \rangle$ are not the same. A general algorithm designed to pick an optimal vocabulary should, if possible and not too costly, take this asymmetry into account.

There is also a basic question about the method by which the confusability measures $\langle a_n^i, a_m^j \rangle$ are formed. Each utterance of $a_n^i$ by each person creates a machine score of $a_n^i$ against $a_m^j$. If we are permitted the luxury of finding an optimal vocabulary for each speaker, we must at least average the machine scores of $a_n^i$ against $a_m^j$ over time to achieve representative and time-stable confusability measures. Since it is more likely that we will wish to choose a vocabulary optimal for several speakers, the machine scores must be averaged over speakers as well. Even with these attempts to render the confusability measures less sensitive to variations over time and between speakers, there is still some question as to whether an optimal vocabulary is stable in time even for a fixed group of speakers. When the speaker group changes, there is even greater uncertainty about the stability of the optimal vocabulary.

# APPENDIX B

# VOCABULARY OPTIMIZATION AS A MULTICRITERIA DECISION PROBLEM

In this memo we will address the problem of determining an optimal vocabulary for a task environment based upon a number of noncommensurable measures of "goodness of vocabulary ". We will do this, first of all, by introducing a general context in which we can meaningfully talk about making decisions when there are a large number of feasible alternative decisions and when the criteria for making these decisions are largely incomparable or noncommensurable. Having done this, we will identify the vocabulary optimization problem as such a decision problem and discuss the solution in a general context. Then we will examine the five criteria for goodness of vocabulary used in this study, identifying the relevant decision problem parameters. Finally we will discuss some drawbacks of this approach and some possible alternative approaches.

## Decision Problems with Multiple Noncommensurable Objectives

In solving many problems relating to system design or system control we are often faced with the problem of optimizing system performance under conflicting multiple performance objectives. For this type of decision problem, a logically sound and universally accepted solution concept does not yet exist. Nevertheless, it seems clear that a good decision should not be dominated by any of the other feasible alternatives, in the sense that there is no other feasible alternative which would yield greater satisfaction to the decision maker.

Suppose that in a typical decision problem of this kind there are N objectives ($N \geq 2$) and that some of these objectives may be conflicting. We will call the degree of satisfaction of an objective the index of that objective, and we will try to maximize the index of each objective. (If the minimization of some index is desired instead, maximize the negative of that index.) Let $\mu_i$ ($i$ = 1, ..., N) denote the index of the $i^{th}$ objective, and call ($\mu_1, \ldots, \mu_N$) the index vector for these multiple objectives. (For example, we could think of $\mu_i$ as the value of the $i^{th}$ real-valued objective function.) Can all these indexes be maximized simultaneously? That is, does there exist an ideal, most desirable, optimal index vector ( $\mu_1^*, \ldots, \mu_N^*$ ) among those attainable such that $\mu_i^* = \max \mu_i$ for i = 1, ..., N? Unfortunately, the answer generally is no, although there do exist cases where the answer is yes and the indexes are not in real conflict.

Let us consider an alternative definition of optimality. Suppose a vector ($\mu_1, \ldots, \mu_N$) is called superior to another vector $\mu' = (\mu_1', \ldots, \mu_N')$ if $\mu_i \geq \mu_i'$ for all i and $\mu_i > \mu_i'$ for at least one $i$. Then an optimal index vector is an attainable index vector to which no attainable index vector is superior. Such an optimal index vector is called

In particular, consider a problem in which the multiple indexes are evaluated according to given multiple objective functions $f_1, \ldots, f_N$ . These functions $f_i$ are to be real valued functions defined on a decision space $X$ so that for each decision $x \in X$, the $i^{th}$ objective function $f_i$ evaluated at the decision $x$ yields the index $\mu_i$ : $f_i(x) = \mu_i$ . A solution to the decision problem is called feasible if it satisfies the constraints of the problem. A feasible solution $x^*$ (a decision) is said to be a Pareto optimal solution if there exists no feasible solution $x$ such that $f_i(x) \geq f_i(x^*)$ for all $i$ and $f_j(x) > f_j(x^*)$ for at least one $j$ . In this case ( $f_1(x), \ldots, f_N(x)$) is an attainable index vector if $x$ is an attainable solution. It is clear from the discussion above that a Pareto optimum can be defined via the notion of a maximal vector, and this turns out to be one effective characterization to use when developing an algorithm to find Pareto optimal solutions.

In a well-posed decision problem of the kind we have been discussing, there always exists at least one maximal vector (and hence a Pareto optimal solution). But maximal vectors are generally not unique. In some problems there may be just a few; in others there may be quite a large number. It may happen that a significant percentage of all attainable solutions are Pareto optimal.

There are several known algorithms whose purpose is to determine the set of all Pareto optimal solutions to a decision problem of the kind described above. One algorithm, designed to work in quite general circumstances, is the PEC (Proper Equality Constraints) method described by J.G. Lin in (1). This method uses a generalized notion of maximal vectors called quasisupremal vectors, and provides a usable technique for generating the class of all such quasisupremal vectors. Hence, practically speaking, this method generates all Pareto optimal solutions of a decision problem. Two other methods, due to Yu (3), generate all Pareto optimal solutions of a decision provided some technical conditions are met regarding the convexity of that subset of N-dimensional Euclidean space consisting of all attainable index vectors. All the algorithms are described in the references cited, so there is no profit in including them here. Other approaches to the problem of finding Pareto optimal solutions in our context are described in Raiffa (2).

### Vocabulary Optimization as a Multicriteria Decision Problem

Let us now look at the vocabulary optimization problem itself and see how this problem can naturally be construed as a multicriteria decision problem. First of all, we consider the general setting of the problem.

We are concerning ourselves with a task environment in which there are a number of functions to be performed. Each function is to be associated with a single utterance consisting of one or more words. We assume that, for each function to be performed, a number of utterances have already been chosen to be associated with that function. (For example, if the task environment is an aircraft cockpit, a function to be performed might be lowering the landing gear, and candidate utterances to be associated with this function might be "wheels down ","lower landing gear ","bring wheels down ", "lower wheels ",and "drop wheels ".) We call the total collection of candidate utterances for all functions the lexicon and a potential vocabulary is a subset of the lexicon which contains precisely one utterance for each function to be performed.

How many potential vocabularies are there? Suppose there are N functions to be performed in the task environment and that the $i^{th}$ function has $n_i$ candidate utterances associated with it. Then the total number N* of possible choices of potential vocabularies is the product of the $n_i$ - that is

$$N^* = \prod_{i=1}^{N} n_i \quad .$$

To get a feeling for the magnitude of N*, let's assume that there are twelve functions to be performed and that each function has four candidate utterances associated with it. Then the total number of possible potential vocabularies is $4^{12}$ = 16 777,216.

It soon becomes clear that, except in cases where N and the $n_i$ are all relatively small, there are too many possible potential vocabularies for us to be able to find an optimal one - in almost any reasonable sense of "optimal" - by exhaustive search.

One reasonable path to optimality through this maze of potential vocabularies is to regard vocabulary optimization as a multicriteria decision problem in the following way. First of all, decide in a general way what a good vocabulary ought to be and set up a number of criteria which individually yield measures of goodness of a potential vocabulary. Secondly, assign an index to each vocabulary for each criterion. Thirdly, in the decision space consisting of all potential vocabularies, determine those potential vocabularies which are Pareto optimal. In a certain sense, we have then solved the problem.

What are the individual criteria which we will use to give measures of goodness of potential vocabularies? In the present study, there are criteria based on five features of the vocabulary items. They are: familiarity of speaker with vocabulary, meaningfulness of the vocabulary, length of utterances in the vocabulary, acoustic confusability of the utterance relative to the human ear, and acoustic confusability of the utterance relative to an automatic speech recognition device. Each of these criteria serves to look at one facet of "goodness of vocabulary," and the overall idea to which they are subordinate is that a good vocabulary is one which is accurately and efficiently used by an operator in a man-machine interface. Furthermore, each criterion is quantifiable, in the sense that a number - an index of satisfaction of that criterion - can be attached to each potential vocabulary.

Decision space, for the vocabulary optimization problem, simply consists of all potential vocabularies. More precisely, a decision X is an N-vector $(X_1,X_2...,X_N)$ where $X_i$ is a candidate utterance for the $i^{th}$ function. So a decision - a point in decision space - is a choice of a vocabulary. We assume that corresponding to each such decision we have a five-vector $(Z_1,Z_2,Z_3,Z_4,Z_5)$, where $Z_j$ is the index of satisfaction of the $j^{th}$ criterion. So, if the first criterion is familiarity of speaker with vocabulary, then $Z_1$ is the index which measures how good this vocabulary is with respect to that criterion. Each of these indexes has to be set up in a consistent way for each criterion, and we will discuss some reasonable ways of doing that in the next section.

Once decision space is set up and the indexes of satisfaction are established, we are in a position to use one of the algorithms to find the set of Pareto optimal vocabularies. That is, we must find those potential vocabularies whose index vectors $(Z_1, Z_2, Z_3, Z_4, Z_5)$ are maximal. Then we are finished.

Getting the Indexes of Satisfaction

An important step in the process of determining an optimal vocabulary using the method we have described above is the establishment of the indexes of satisfaction. This is primarily an issue of creating in a reasonable and consistent way a quantitative measure for each criterion of "goodness of vocabulary" that we have chosen, and, secondarily, a matter of using these quantitative measures intelligently.

For each of the non-acoustic criteria used in this study – that is, for length, meaningfulness, and familiarity of utterance – a real number can be attached to each utterance in the vocabulary in such a way that if measure $M_1$ is attached to vocabulary item 1 and $M_2$ is attached to vocabulary item 2 and if $M_1 < M_2$ the second vocabulary item is preferable to the first for the criterion that the measures $M_1$ and $M_2$ are measuring. So a single real number can be attached to each utterance in the lexicon. In this case, no matter how the quantification actually takes place, it is possible to assign an index to each potential vocabulary for each non-acoustic criterion. For example, if the items in an N-utterance vocabulary have individual measures $\nu_1$, $\nu_2$, ...., $\nu_N$, an index $\mu$ for that whole vocabulary could be defined by

$$\mu = \sum_{i=1}^{N} \nu_i$$

That is the index $\mu$ is just the sum of the "measures of goodness" of the individual vocabulary items.

For the two acoustic criteria - acoustic confusability relative to human and relative to machine - the definition of an index is not so simple. The main reason for this is that acoustic confusability is inherently a property of <u>pairs</u> of utterances, and not of single utterances by themselves. So our definition of the acoustic indexes is going to be a bit more complicated.

For a potential vocabulary $V = \{V_1, V_2, \ldots, V_n\}$ let $C_{ij}$ be the confusability measure of $V_i$ with respect to $V_j$. This confusability measure is a score indicating acoustic similarity between the item $V_i$ and the item $V_j$. There is one scoring system for human acoustic confusability, and another one for machine acoustic confusability. Under some scoring schemes it is possible that the scoring is not symmetric - that is, $C_{ij} \neq C_{ji}$. We will symmetrize the confusion matrix $\mathcal{C} = (C_{ij})$ by replacing $\mathcal{C}$ by $\underline{\mathcal{C}} = (\underline{C}_{ij})$ where $\underline{C}_{ij} = \underline{C}_{ji} = \max(C_{ij}, C_{ji})$. We then establish a so-called lexicographic ordering on the class or all potential vocabularies as follows.

First, for each vocabulary, list the confusability measures between each pair of items in order with the largest (most confusable) first, and the smallest (least confusable) last. Suppose $V$ and $V'$ are two vocabularies, and suppose we have ordered their confusability measures $\{\mu_n\}$ and $\{\mu_n'\}$as we described above so we have

$$\text{For } V: \quad \mu_1, \mu_2, \mu_3, \ldots, \mu_{N(N+1)/2}$$

$$\text{For } V': \quad \mu_1', \mu_2', \mu_3', \ldots, \mu'_{N(N+1)/2}$$

We will say that $V'$ is greater than or equal to V, and write $V' \geqslant V$ if <u>either</u>
(1) $\mu_1' < \mu_1$ , <u>or</u> $\mu_1' = \mu_1$ and $\mu_2' < \mu_2$ <u>or</u> $\mu_2' = \mu_2$ and $\mu_3' < \mu_3$ ...etc., <u>or</u>
(2) $\mu_i' = \mu_i$ for all i.

The relation $\geqslant$ puts a linear order on the class of all potential vocabularies, and allows us to assign indexes to vocabularies starting with zero for the worst (most confusable) vocabulary, and increasing by one for each succeeding vocabulary but assigning equal indexes to vocabularies with $\mu_i' = \mu_i$ for all i. This procedure thus allows us to assign indexes of satisfaction for both kinds of acoustic confusability.

<u>Some Drawbacks of Pareto Optimization and Some Alternatives</u>

The principal disadvantage involved in applying the techniques of Pareto optimization to the vocabulary optimization problem is that we are likely to get too many solutions. When a substantial percentage of all potential vocabularies turn out to be Pareto optimal, as is likely here, it is difficult to feel that we have made much progress in drawing meaningful distinctions between the vocabularies. Furthermore, all the criteria - all measures of "goodness" of vocabularies - are treated equally, as is required in all the standard approaches to Pareto optimization. (We can see that there is a reasonable and consistent generalization of Pareto optimality that does allow us to put a partial ordering on the measures of "goodness" of vocabularies, thus creating a ranking of the criteria for judgment. But we were not able to find this approach followed up anywhere in the literature of the subject.)

The main advantage to the use of Pareto optimization techniques is that they are the only techniques which provide a logical and consistent solution of the problem as stated. If we modify the conditions of the problem somewhat, then some other solutions are possible. One solution - a shaky one at best - is to establish a weighting scheme for assigning weights to the indexes associated with each criterion. The overriding problem with this approach lies in the selection of the weighting functions. In this study, there is no compelling reason at all to pick any one weighting over any other since the inter-relations between the criteria of judgment are simply not known. Of course, if these inter-relations were known, the whole problem would be a lot simpler and Pareta optimization techniques would be inappropriate.

Another alternate solution to the problem of finding an optimal vocabulary is to modify the notion of optimality that we are using. Suppose, for example, that we decide that one criterion for "goodness of vocabulary" is significantly more important than the other criteria. Then we can use the other criteria simply to eliminate bad vocabularies, and then optimize among the remaining vocabularies using only that most significant criterion. In the case that the most significant criterion is acoustic confusability relative to machine, we could use the other criteria to eliminate those vocabularies which are particularly bad with respect to length of utterances, familiarity, etc. and optimize on the reduced class of vocabularies. This seems to be a viable technique here because picking an optimal acoustically non-confusable vocabulary can be accomplished using a tree search algorithm with automatic pruning of the tree at each stage of the algorithm. For a large class of vocabularies to consider, this algorithm might take a long time to find an optimal vocabulary but the other, less significant, criteria could be used to prune the initial class of vocabularies significantly. The principal disadvantage of this technique is that it uses all but one of the basic criteria only negatively to rule out possible vocabularies and does not make use of the potential abilities of these criteria to make finer discriminations.

## Summary and Conclusions

In this memo we have presented an approach to vocabulary optimization based upon viewing this optimization as a multi-criteria decision problem. We described the process of setting up the indexes of satisfaction for the decision problem based on several criteria, and we also suggested that Pareto optimality gives a viable, although by no means ideal, notion of overall "goodness of vocabulary ". We noted that the number of Pareto optimal vocabularies is likely to be large indicating that this notion of optimality does not make very fine distinctions among vocabularies. However, the technique of Pareto optimization does produce a number of good vocabularies which could reasonably be examined further experimentally to determine a smaller class of vocabularies optimal under specific operating circumstances.

## REFERENCES

(1) Liu, J.G., "Multiple Objective Optimization: Proper Equality Constraints (PEC) and Maximization of Index Vectors" in Multicriteria Decision Making and Differential Games, edited by G. Leitman, Plenum Press, New York, 1976

(2) Raiffa, H., "Preferences for Multi-Attributed Alternatives," The RAND Corporation, Memorandum No. RM-5868-DOT/RC, 1969.

(3) Yu, P.L., "Cone Convexity, Cone Extreme Points, and Nondominated Solutions in Decision Problems with Multiobjectives," in Multicriteria Decision Making and Differential Games,edited by G. Leitmann, Plenum Press, New York, 1976.

## APPENDIX C

## VOCABULARY SUGGESTIONS FOR IMPLEMENTING ASR IN AN F-4 COCKPIT

Listed below are the preliminary vocabulary suggestions for implementing ASR in an F-4 cockpit to control the following functions:

| | | |
|---|---|---|
| a. | Emergency relese of external stores | JETTISION ALL |
| b. | Jettison the centerline station | JETTISON CENTERLINE |
| c. | Missile control: System selection | SELECT PLM<br>SELECT VTAS |
| d. | Deploy the drag chute | DRAG CHUTE |
| e. | Relase the hook | HOOK DOWN |
| f. | Selective missile jettison | JETTISON STATION X |
| g. | Set gunsight | GUNSIGHT XXX MILS |
| h. | Lower (Raise) the landing gear | WHEELS DOWN<br>WHEELS UP |
| i. | Probe control: Refuel all tanks | EXTEND PROBE: ALL TANKS |
| j. | Probe control: Refuel feed tanks first | EXTEND PROBE: FEED TANKS |
| k. | RAT control | EXTEND RAT<br>STOW RAT |
| l. | Jettison the external wing tanks | JETTISON WING TANKS |
| m. | Radio: Frequency (Channel) control | SWITCH TO XXXX |
| n. | Radio: ADF: Frequency (Channel) | ADF XXX.X |

| | | |
|---|---|---|
| o. | Change TACAN channel | TACAN CHANNEL XXX |
| p. | Radar control | RADAR: Range XXX<br>pulse<br>low<br>high |
| q. | Select IFF code | SQUAWK IDENT |
| r. | IFF: System selection: Secure | IFF OFF |
| s. | Couple the data link | COUPLE |
| t. | Altimeter setting | SET ALTIMETER XXX |
| u. | Release flares or chaff | FLARES (CHAFF): BURST (SALVO) |

DISTRIBUTION

| Addressee | DODAAD Code | No. of Copies |
|---|---|---|
| Scientific Officer<br>Director, Electromagnetics Technology Program<br>Assistant Chief for Technology<br>Office of Naval Research<br>800 North Quincy Street<br>Arlington, Virginia 22217<br>Attn: CDR D. C. Hanson | N00014 | 1 |
| Defense Contract Administration Services<br>Management Area, San Diego, Bldg. 4,<br>AF Plant 19, 4297 Pacific Highway<br>San Diego, California 92110<br>Attn: DCRL-GSCA-73 | S0514A | 1 |
| Director, Naval Research Laboratory<br>Attn: Code 2627<br>Washington, D.C. 20375 | N00173 | 6 |
| Defense Documentation Center<br>Bldg. 5, Cameron Station<br>Alexandria, Virginia 22314 | S47031 | 12 |
| Office of Naval Research Branch<br>Office - Pasadena<br>1030 E. Green Street<br>Pasadena, California 91106 | N62887 | 1 |