DOT/FAA/AM-96/10
Office of Aviation Medicine
Washington, D.C. 20591

Methods and Metrics of Voice Communications

Barbara G. Kanki, Editor
Ames Research Center
National Aeronautics and Space Administration
Moffett Field, CA 94035

O. Veronika Prinzo, Co-Editor
Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma 73125

Proceedings of a workshop organized by the Federal Aviation Administration, National Aeronautics and Space Administration, and Department of Defense. San Antonio, Texas, May 13-14, 1995.

March 1996

Final Report

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

U.S. Department
of Transportation
Federal Aviation Administration

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-96/10		
4. Title and Subtitle Methods and Metrics of Voice Communications		5. Report Date March 1996
7. Author(s) B.G. Kanki and O.V. Prinzo		8. Performing Organization Report No.
9. Performing Organization Name and Address FAA Civil Aeromedical Institute P.O. Box 25082 Oklahoma City, OK 73125 NASA Ames Research Center Moffett Field, CA 94035-1000 USAF Office of Scientific Research Armstrong Laboratory Brooks Air Force Base, TX		10. Work Unit No. (TRAIS) 11. Contract or Grant No. DTFA-02-91-C-91089
12. Sponsoring Agency name and Address Office of Aviation Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, D.C. 20591		13. Type of Report and Period Covered 14. Sponsoring Agency Code

15. Supplemental Notes

This work was performed under Task AM-D-94-HRR-117

16. Abstract

This report consists of the proceedings of the Methods and Metrics of Voice Communication Workshop organized by the FAA - Civil Aeromedical Institute, NASA-Ames Research Center, and Armstrong Laboratory - Brooks Air Force Base, held May 13-14, 1994 in San Antonio, Texas. The goal of the meeting was to further our understanding of voice communications in aviation operations by convening a group of language researchers to discuss their experiences with current methods, tools, approaches, etc. The participants came from academic, government and military laboratories, as well as private industry, and their collective expertise included applications conducted in a variety of field, laboratory and simulation environments. This Proceedings is presented in 3 parts: (1) presentations on discourse and acoustic processes, (2) 'demonstrations' of software aids for collecting, coding and analyzing communication data, and (3) an appendix of related supplementary materials and reprints from other publications. Because language researchers share many issues and problems in the development of appropriate and effective methods and metrics, the goal of the workshop was for the attendees to learn from each other's experiences and expertise and to make these insights available to others.

17. Key Words	18. Distribution Statement
ATC communications Discourse processes Acoustic processes	This document is available to the public through the National Technical Information Service, Springfield, VA 22161.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	285	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

PREFACE

The realm of language research is populated with a large number of diverse data collection, transcription, and analytic techniques from which language researchers select approaches based upon their particular set of individual research objectives. To date, this methodology has not been assimilated in a useful form to facilitate its use by the language research community. In an attempt to achieve some form of assimilation, Dr. O. Veronika Prinzo, Federal Aviation Administration (FAA), Dr. Barbara G. Kanki, National Aeronautics Space Administration (NASA), and Dr. Samuel G. Schiflett, United States Air Force (USAF), conceived a jointly-sponsored symposium to gather together a body of experts in the field of voice communications and attempt to further the collective understanding of the methods and metrics used in the study of natural language.

The workshop concentrated on a diverse collection of techniques and approaches used to analyze both discourse and acoustic processes in voice communications. Discussions focused on data collected from simulation/laboratory environments, as well as from field and case study investigations. Issues included (1) determining units of analysis, (2) coding and statistical techniques, (3) approaches to filtering the speech signal, (4) strategies for integrating verbal and non-verbal communications, (5) data collection and research design issues, and (6) software applications.

During 2 days of presentations and demonstrations, the participants (Figure 1) shared past experiences and research findings, current interests and information, as well as future plans and opportunities. This document reports the information as it was presented at the workshop and also provides a resource for other language researchers.

FIGURE 1: Methods and Metrics of Voice Communication Workshop Participants.
Top Row: Malcolm Brenner, Leon Segal, David Pisoni, Clint Bowers, Lawrence Porter, Alan Reich.
Row 3: Doug Eddy, Penny Sanderson, Joe Danks, Lynn Nygaard, Steve Veronneau, Martin Thee.
Row 2: Herb Clark, Beth Veinott, Dan Morrow, David Mayer, Roni Prinzo.
Row 1: Sam Schiflett, Ann Bradlow, Judith Burki-Cohen, Carol Symer, Barb Kanki.
Not Present: Don Foss, Jeff Whitmore, Linda Connell, Carolyn Prince, Howard Harris, Cheryl Irwin.
Photographer: Linda Barrett.

TABLE OF CONTENTS

INTRODUCTION

PART 3: APPENDIXES

METHODS AND METRICS OF VOICE COMMUNICATIONS

WELCOME AND OPENING REMARKS

I'm Dr. Roni Prinzo, with the FAA's Civil Aeromedical Institute (CAMI). Before we begin, I'll give you a brief history on how CAMI, NASA Ames, and the Armstrong Laboratory at Brooks Air Force Base came together to sponsor and host the Methods and Metrics of Voice Communications Workshop.

When I first came to work for CAMI, the major thing I knew about aviation was how to purchase a ticket and board an airplane. Although I had a doctoral degree in psychology with an emphasis in psycholinguistics, I had no experience in air traffic control (ATC)/pilot communications. Upon listening to my first audio tape of pilots speaking to an air traffic controller, I did not understand what they were talking about. For example, "Regional Approach, roger, out of sixteen for ten with Charlie" held no meaning. To gain an understanding of aviation terminology and local jargon, I read the existing literature, visited several air traffic control facilities, and asked lots of questions as part of my self-directed education in operational communications. Soon I learned that other researchers also had experienced the same or similar problems with aviation terminology. The phrase "communication error" was particularly problematic. Within aviation, it is often used to refer to loss of separation minima by which aircraft are spaced to achieve safe and orderly flight that is attributed to communication. To communication researchers, "communication error" is generally viewed more broadly as any occasion when actions taken are based on faulty communication. Through discussions with different individuals, it became exceedingly clear that a need existed to bring together a group of scientists and professionals interested in communications to share with one another their experiences in communication-based research and to develop some common definitions.

I discussed my perceptions with Dr. Barbara Kanki and several other people from the aviation community. At the 1993 Ohio State Symposium in Aviation Psychology, Barbara and I decided to jointly sponsor a workshop. While at a briefing at Brooks AFB that Fall, I mentioned that a workshop was in the works. Dr. Sam Schiflett commented that he had made a provision to host a similar workshop in his 1995 program. We discussed the possibility of having him become active in the 1994 workshop. Upon my return to CAMI, Barbara and I agreed that Sam should become an integral part of our venture. Then, San Antonio was selected as the site of the workshop.

I'm really delighted that so many of you were able to attend.

THE ROLE OF COMMUNICATIONS IN TEAM SITUATIONAL AWARENESS: A SCIENTIFIC PROGRAM OVERVIEW

Samuel G. Schiflett, Ph.D.
Armstrong Laboratory, Brooks AFB

INTRODUCTION

All branches of the military employ both ground and airborne operational personnel in Command, Control, and Communications (C^3). The mission element of the Air Force in C^3 systems is to provide surveillance, identification, warning, and control in support of tactical and global air operations. A key to the success of all Air Force missions is the rapid establishment and maintenance of distributed communication networks so that accurate and timely information can be exchanged within teams and between our war fighting units. More often than not, the early phases of any war are won or lost on how effectively military personnel communicate with each other. It forms the basis of all tactical, strategic, and intelligence coordination activities whether it is within a flight crew, between mission elements, or at centralized command headquarters. Communications is simply the technical means to achieve control, and control is simply the structural means to command.

Technological advancements in high-speed, wide-band communication networks have greatly diminished the restrictions imposed by past communications systems in volume, rate, and type of information transmitted (Vincent, 1993). The advent of orbital communication satellites coupled with precise global positioning networks have expanded the capability of an individual with a single hand-held communications transmitter and receiver to exchange information with a myriad number of world-wide data links. Improvements in the design of these more flexible communication systems offer inter-operative links to remote operational units that previously could not exchange information in an efficient manner. Even though these new communication systems have brought many enhancements to cooperative planing and engagement phases of Air Force missions, they have also imposed a greater need for team and unit coordination. Technology has by far outpaced our understanding of how the newly acquired information should be presented, what hierarchical level should receive it and act on it, and what effect the increased alternatives will have on decision making and team performance. More information does not necessarily equivocate to better TEAM Situational Awareness.

BACKGROUND

Individual Air Force personnel most often perform their jobs as part of a team. Military team performance affects such diverse functions as command and control, flight and ground crew tasks, acquisition, design, maintenance, logistics, and others. However, the impact of team performance cannot be evaluated or even measured until the task demands are identified and thoroughly described. What common features do these seemingly diverse jobs impose on team performance? How do people make decisions in situations characterized by these features? Some of the more salient characteristics of C^3 systems described by Rouse, Cannon-Bowers, & Salas (1992) and Orasanu & Salas (1993) include the following:

1. Team (crew) members are composed of individuals that have been assembled to complete a required task (mission). Consequently, individual decisions and actions must be viewed in the context of accomplishing a team goal.
2. There is no single predetermined solution to a problem. Team conformity to standard operating procedures (plans) should be discarded or modified if individual strategies provide a more accurate and timely solution.
3. Members of a team have specialized knowledge and skills relevant to the decision and overall task assignment. Therefore, team communication and coordination are central issues in distributed decision-making research.
4. The work situation is highly dynamic (changing priorities and varying tempo) and externally driven. Autonomous teams must frequently adapt to changing circumstances by making decisions for others under time constraints.
5. Individual and team quality of performance have significant consequences. Team members often make decisions and take actions that will place them or others at risk.

Unfortunately, classical decision-making research has not offered useful explanations of how teams function, as characterized by these situations. Despite thousands of studies and large scale military support for behavioral decision research involving Bayesian statistics, Klein (1993) has observed that overall the results have been disappointing. One of the reasons is these models have focused on highly structured, predefined tasks where there were only correct and incorrect binary decisions. While the models have been useful for studying college sophomores performing context-free laboratory tasks, they hold little relevance for the complexities of command-and-control settings as described above. A shift in the domain of basic research is necessary if the role of communications in team situational awareness is to be understood and explained in distributed decision-making environments.

The research domain selected for this scientific program, sponsored by the Air Force Office of Scientific Research, has been formulated out of an operational need to improve situational awareness within and between teams in complex decision-making environments. This program overview describes, in more detail, the research domain, scientific goal, sub-goals, approach, research paradigm, objectives, measurement methodology, and communication measures.

Research Domain

Basic research will be conducted to critically examine theories and empirically verify derived postulates that can relate the dynamics of communication to the formulation of shared models of complex changing environments. This research initiative will develop a measurement methodology to study the sharing of information between team members in support of individual actions and group success. Specific research issues concern the nature of cognitive models held by individual team members that support effective communication; the information requirements of individuals that support coordinated behaviors; the internal models that members have of each other that affect the quantity and quality of information transmitted; the relationship between infrastructure of a team and its ability to function effectively under specific task demands; and the determination of types of teams based on differences in their behavioral characteristics, e.g. content and pattern of interactive communications.

Scientific Goal and Sub-goals

The scientific goal is to initiate a long-term research program that fosters scientific collaborations focusing on the underlying mechanisms of team performance to gain an understanding of the role of communications in enhancing and maintaining situational awareness in distributed team decision-making.

The scientific sub-goals are to develop techniques of measurement of team communication and verify the concept of shared mental models in more complex and stressful environments.

APPROACH

A set of interrelated team constructs (coordination, conformity, cohesiveness, composition, and adaptability) will be defined and propositions will be presented to empirically verify the interactive role of communication variables in explaining and predicting the effect on situational awareness.

The formal structure of coordination in teams will be analyzed by specifying input, process, and outcome variables that affect a team member's decisions to communicate with information sources to accomplish task-specific assignments. Of particular interest is the degree to which one team member has the same situational understanding (shared meaning) of the significant events, current status, and future projections in relationship to the other members. It is interesting to note that all or part of the team members could have a shared perspective that differs from the actual situation. A definition of some of the constructs and boundaries of research will help elucidate the study objectives.

Definitions

There are as many definitions of what a team is as there are researchers trying to define a team. The essential difference between teams and other problem-solving groups with common goals is the nature of the tasks they face and the behavioral responses required for their completion. The most applicable characteristics of teams to this research is defined by Dyer (1984) and discussed by Morgan, Glickman, Woodard, Blaiwes, and Salas (1986). The essential elements are as follows:

> A **team** consists of "a distinguishable set of 2 (3) or more people who interact interdependently and adaptively to achieve specified, shared, and valued goals (mission objectives)." (Morgan et. al., 1986 p.3)

The parenthetical inserts shown in the above definition of a team were modified by this researcher to emphasize more of a military command and control working environment. An additional emphasis by this researcher on the size of the lower boundary of a team excludes dyads. Dyads (e.g., pilot & co-pilot) are often considered a team. However, they are excluded from this research if they are studied in isolation, because there are a number of important team processes that do not occur in only two-person interactions. For example, coalition formation, complex patterns of status, and more importantly to this research, hierarchical communication patterns (Ilgen, Major, Hollenbeck, & Sego, 1991).

Situational Awareness (SA). The more operational definition of situational awareness from a pilot's perspective is "a continuous perception of self and aircraft in relation to the dynamic environment of flight, threats, and mission, and then to forecast, then execute tasks based on that perception." This definition was a consensus statement forged together by the Situational Awareness Integration Team (SAINT) assembled to plan and conduct an integrated program of research to develop and validate situational awareness measures. The SAINT team was directly commissioned by General Merrill A. McPeak who supplied his own definition of situational awareness shortly after the completion of DESERT STORM, "I know it ... when I see it." That was interpreted by the Air Combat Command to mean "the capability to appropriately assess yourself, your system, and your environment in order to make the right decision and the right response at the right time." Good situational awareness is part of having the "Right Stuff" as popularized in today's fighter pilot jargon.

Perhaps, a more scientifically acceptable definition of situational awareness for basic research is offered as follows:

> The (identification) perception of elements (events) in the environment within a volume of space and (a stream of) time, the (shared) comprehension of their meaning, and the projection of their status into the near future. (Endsley, 1988)

Again, a few parenthetical modifications by this researcher warrant an explanation. The perception, comprehension, and projection process that Endsley describes seem to be exactly what a pilot or weapons controller means when they say you must "stay ahead of the game" to be successful. The words and phrases inserted into the definition are an attempt to expand the statement into more of a team definition of situational awareness. Team members should possess a common (shared) understanding of the nature of events impacting others. The identification of events in a "stream of time" emphasizes that a ground-based or an air weapons controller has a different perspective of the rate of change from a fixed reference point as compared to a pilot traveling at high-speed at low levels. The pilot is rapidly moving through a "volume of space" which results in a perception of the events as being time compressed.

Conversely, the AWACS air weapons controller loitering at 40,000 feet is in a "stream of time" with emerging events of unequal priorities and diverse time constraints. The air weapons controller must maintain an accurate "big picture" of the battle because this defines the awareness of the current situation in relationship to past and future events. Through situational awareness, the air weapons controller chooses among the tasks competing for attention and then executes the most important. This decision-making process is more than the application of a predetermined set of individual priorities. As Dalrymple (1991) has emphasized, the choices should be from a team member's perspective in determining what trade-offs will increase options in the future, what will ease future workload, and what will be the most expedient to implement.

Taylor (1989) offers an empirically based definition of situational awareness taken from a factor analytic approach of interviewing aircrew of how SA is actually experienced and what elements comprise it. One of the dominant factors was the construct of understanding in the form of information quality, quantity, and familiarity. These features of information will be manipulated in this research program to obtain a better understanding of the "perception of elements" commonly held by the team members.

Perhaps one of the most parsimonious definitions of team situational awareness is the "collective knowledge needed to sustain adaptive coordinated behavior in a changing environment necessary for survival and mission success." This definition, supplied by Dr. John Tangney, the AFOSR Program Manager, emphasizes: (1) Types of knowledge of other members, environment, and task domain; (2) Residence of knowledge that is either shared or distributed, human or computer; and (3) Communications needed to build and sustain knowledge, perform optimally, and recover from errors. This definition of team SA has the added benefit of introducing an essential team construct of adaptability.

Research Paradigm

The proposed experiments will be conducted in a controlled setting that will allow the examination of constructs that are important to team coordination in operational environments. The essential difference between this type of team research and other problem-solving group experimentation is the nature of the tasks they face and the behavioral responses required for their completion. The team tasks that will be used in this research program will consist of specialized component tasks that require coordinated responses to achieve maximum team performance. The research paradigm will focus on *adaptive team coordination* in a wide range of scenarios and conditions.

The general procedure for investigating adaptive coordination in teams will use "scripted" events embedded into realistic scenarios that alter the task demands in specific ways to test theoretical constructs. The changes in task structure and dynamics will require the team to adapt their strategy to form new patterns of communication in seeking and transmitting information. How does the team adapt to the changed situation? How long will the team attempt to follow a plan that is not working? How will the team members communicate the information to each other? The task structure will correspond to that which is actually faced in an operational mission. For example, in a Defensive Counter Air mission, the primary goal of an air weapons controller team on-board an

AWACS aircraft is to "protect friendly assets." The goal can be subdivided, according to Dalrymple (1991), into specific objectives that require a team of operators to detect, identify, intercept, and destroy hostile aircraft in a particular zone of air space. Each team member has an area of assigned responsibility. The computer-generated hostile aircraft symbology can be "scripted" to follow predetermined tracks to weave in and out of each member's protective air space to evaluate team coordination activities. This embedded event will elicit a sequence of behaviors that are directly measurable in the form of information exchange in verbal communication networks.

Synthetic tasks. Another approach this research paradigm will emphasize in studying adaptive coordination in teams is to develop "synthetic tasks" that functionally represent the higher-fidelity scenarios. Lower-fidelity tasks (synthetic) will be created and arranged to present the same time constants of interaction with the environment as the higher-fidelity scenarios. The modular synthetic tasks will offer incentives for individual and team payoffs that reflect the costs (loss of resources) and rewards (number of hostile strike completions) of modern command and control wargames. For example, the team payoff will be higher if the individual members take into account both their own situations and the likely actions of their team mates. This will allow investigations into concept formation of shared mental models of team situational awareness based on a set of known contingencies. The term "mental model" refers to the cognitive representation each team member has of the team scenario, including the team goal, task strategies, and current status of performance and anticipated future requirements (Swezey & Salas, 1992). The higher the overlap in team member shared situational awareness, the higher the expectation team members will have of accurate representations of the needs of the other team members to make effective decisions. The findings from these controlled laboratory experiments will be compared to the results of the "bench mark" higher-fidelity scenarios to verify the theoretical constructs and underlying mechanisms of verbal communications.

Distributed Decision-Making Network. The "synthetic tasks" will be computer-based and therefore need not be run face-to-face nor have all members of the team in the same location. In the later phases of this scientific program it is planned that university and government laboratories will be connected to a wide-area network to conduct collaborative studies in distributed decision making. The objective of this phase of the program is to better understand the role of communications in formulating and maintaining situational awareness when it is carried out in a distributed decision-making environment.

While there may be task conditions that enable a geographically distributed team to outperform a contiguous team, our present view of the consequences of physically separating team members will be almost entirely negative. For example, it is anticipated, based on our current experience in conducting collaborative high-fidelity Defensive Counter Air missions with the Human Resources Directorate at Williams AFB, AZ, that team situational awareness will be diminished because key information will not enter the collective knowledge base of dispersed team members. Teams are part of a larger organizational hierarchy that have a different framing reference (culture) in deciding what should be verbalized and what is understood. Also, if the team has been trained face-to-face and is then geographically dispersed, the loss of a node (unique source of information) in the network can severely disrupt communications that are required for shared team situational awareness.

Klein and Thordsen (1990) have found that team decision-making in many ways resembles individual strategies but there are emergent problems and dysfunctions that can only appear in a distributed team context. For example, poor communication of cues and events transmitted to other team members can result in the total loss or degradation of information critical in maintaining adaptive team coordination. Thus, distributed decision-making networks will be used to examine how remote team members improvise (change strategies), under conditions of information uncertainty, to reach the mission objective.

SCIENTIFIC STUDY OBJECTIVES

The focus of this research is on the role of communications in the formulation and maintenance of team situational awareness. The central theme of the scientific study objectives is the quality and timeliness of exchange of information between team members. The research is restricted to the potential effects on the loss or degradation of information supplied by other team members that is necessary to carry out an individual task in relationship to the team goal. That is, we are not interested in individual task specific performance but how loss or degradation of information impacts the team as a whole.

Information may be either completely lost by total external communication failure (message sent but not received) or denied from an originating source (message not sent). More than likely, in the real world, the information is either degraded by physical noise e.g. partially-jammed communication links, or corrupted unintentionally by passing partially-correct information to other team members. However, the consequences might be quite different depending on the level of collective awareness that each team member has that an error has been introduced. If communications are being physically jammed, the information becomes immediately suspect and a mental model is formed by team members that an unreliable database currently exists for assessing other team members' circumstances. However, subtle incorrect information circulating in the communications system either introduced by misinformed team members or misinterpreted by other team members may perpetuate false ideas and concepts which lead to poor decisions by others. The effects of the loss or degradation of information from others depend on the extent of use of that information by the recipient and whether functionally redundant information is available. The following study objectives will investigate some of these issues.

Study Objective 1

Study objective 1 is to evaluate the effects of the loss and degradation of *redundant* and *non-redundant* information in adaptive team coordinated behavior. How do team members who have achieved a high degree of coordination interpret their situations, change their behavior, and realign their sub-goals when the information necessary to maintain team performance is no longer available or not reliable enough to be trusted? It is hypothesized that only the loss or degradation of non-redundant information that was actively used in maintaining coordination should affect coordination and interfere with team performance. If redundant information is available, team coordination will be regained once the team member with good situational awareness adapts to its use. Redundant and non-redundant information will be presented in both visual and auditory sensory modes so cross-modality features can be studied.

Study Objective 2

Study objective 2 is to evaluate the team member's perceived reasons for the loss or degradation of the information in adaptive team coordinated behavior. How do team members adapt their coordinated behaviors based on verbal communications when the *origin* of information loss or degradation is from either external sources e.g. jammed communications, or from misinformed team members at different levels of the command hierarchy? It is hypothesized that detection, interpretation, and adjustment to loss or degradation of information will be influenced by each team member's shared mental model of the perceived control of the source of error and the differential status of the team members. External sources of corrupted information will produce less conflict in team coordinated behavior than those attributed to human sources of error. Team composition and familiarity will be a major independent variables.

Study Objective 3

Study objective 3 is to evaluate the effects of communication *architectures* (structure) on the formation and maintenance of shared mental models of team situational awareness. What is the relationship between the communication infrastructure of a team and its ability to function effectively under changing task demands and environments? It is hypothesized that teams with flexible, face-to-face communication architectures derived by team members will perform more effectively than teams that have members geographically separated by a structured distributed decision-making network. The predictive validity of shared mental models will be tested by measuring both the process and outcome measures of team performance in tasks that demand team members to adapt to new domains of situational awareness. Incidental learning of common cues and actions will be contrasted with teams void of such information. Patterns of communications and content of information transmitted that were utilized in successful and dysfunctional team problem-solving strategies will be identified and transitioned to the training and selection community of researchers for further study.

METHODOLOGY

Studies of team performance have unique methodological issues. The paramount challenge in understanding the underlying mechanisms of communications in constructing and maintaining team situational awareness is *measurement*. What to measure is just as important as how to measure. What is the nature of the mechanism that allows members to work together in a team tasking situation where interdependence is a key to a successful mission outcome? Individual communication process variables are the most outward and measurable manifestation of team interaction. Recognizing that verbal communication is a vital mediator of information, brings a better understanding of the type of measures required to explore the underlying dynamics that influence team coordination and affect team performance.

It has been observed by Foushee (1984) that most research studies of team performance have ignored communications as a process variable. Past investigations have generally concentrated upon direct links between team input variables (size, structure, composition) and performance output variables (quality, latency, errors). This approach to team measurement may be the prime contributing reason why the literature on team performance exhibits so much inconsistency. Examination of communication patterns and content of speech as a team process variable will often indicate that they are moderating the relationship between input and output variables.

For example, Siegel and Federman (1973) reported using an analytical framework for coding crew communications by combining the Bales (1950) interaction process analysis methodology and the Osgood semantic differential technique. In the initial study, involving Navy helicopter crews, the authors obtained approximately 30 communication variables; but the content analysis focused on the 14 that related to crew performance outcome variables (e.g., number and distance of targets missed). Factor analysis of these communication moderator variables yielded four factors labeled and described as follows:

Probabilistic Structure: Communications in which event occurrence and risk assessment was discussed; reflective communications containing thought processes which involved the weighing of alternatives and the searching for answers to unresolved questions.

Evaluative Interchange: Communications which contained direct requests for information and opinion, as well as the responses to these requests.

Hypothesis Formulation: Communications involving interpretations of past performance in the mission and the evaluation of future tactics to follow.

Leadership Control: Communications marked by a role-coordinating attitude by the team leader, an attitude that served to define goals and to set a proper atmosphere for effective employment of the other 3 factors.

In the second phase of the study, Anti-Submarine Warfare crews received communications training. Simulator data indicated that the trained group performed better (number of correct attacks) than the control group, without loss of time and navigational accuracy. The relative frequency of all the communication factor categories differed. However, the only statistical significant differences were: (1) the probalistic structure constituting 22% of the communications with the trained group and 11% within the control group and (2) the leadership control category being 41% in the trained group and 60% in the control group (untrained). For the trained group, leadership control meant encouraging an interchange of opinion and information; for the control group it reflected a tighter and more autocratic leadership structure. The authors hypothesized that the differences in communication between the 2 groups may have accounted for the differences in the output variable of crew performance. Thus, communication moderated the outcome.

One of the most significant team process variables reflected in communications is information flow between members of the team. The measurement of relational communications has been utilized over many years by a large number of researchers in various group processing paradigms (e.g. Bales, 1950; McGrath, 1984). As noted by Foushee & Helmreich (1989), in those studies that have examined the relationship between group process variables and performance effectiveness by closely examining group member communications, they have often proven fruitful. For example, Foushee & Manos (1981) analyzed the cockpit voice recordings from the Ruffell Smith (1979) simulation study utilizing a technique adapted from Bales' interaction process analysis. Several interesting relationships emerged from the Foushee and Manos study. Overall, there was a tendency for crews who communicated less not to perform as well, but the type or quality of communication played an even more pivotal role. Perhaps, the most salient aspect of the results of this flight simulation study was the finding that most problems were related to breakdowns in crew coordination, not a lack of technical knowledge and skill. The "high-error" crews experienced more difficulties in the areas of communication style and relevancy of information transmitted than "low-error" crews.

It should be noted that this finding of the source of the communication breakdowns would have been missed if flow (amount) of information (expressed in bits) would have been the principal metric. As originally envisioned by Shannon and Weaver (1964), information theory is not concerned with either the *meaning* or *effect* of the message. Likewise, the actual *content* of the message is unimportant to the measurement of information gain. The important concept in this meaning of information is the set of possible messages that could have been transmitted per unit of time (channel capacity) and the actual number of received messages from this set of equal alternatives. Information is gained only when a message is delivered which reduces uncertainty. Information in this usage must not be confused with the more common definition of "meaningful knowledge."

However, information theory has been a great asset to researchers investigating both communication processes and operator performance. Wickens (1984) suggests that information theory provides an essentially dimensionless unit of performance across a wide variety of different dependent variables. Fitts and Posner (1967) have also suggested that certain limits of the human information processing system remain relatively invariant when described in the terms of information theory. Despite these successes, the use of information theory in human performance research and applications has received some criticism. Among the limitations sighted are insensitivity of the information metric, requirements for structured tasks, and the inability to describe the factors influencing response time. However, as DaPolito, Jones, & Hottman (1989) point out, the utility of information theory in studying team performance is, that it can: (1) Serve as a model for perceptual processes and, (2) Provide a means of evaluating new communications technology by comparing transmission rates (throughput) of information within or between different sensory modalities.

A recently completed study by Hottman, DaPolito, Dalrymple, & McKinley (in press) determined the amount of information and its importance for task completion of an AWACS C^3 mission using a novel methodology based upon a combination of information theory, task analyses, and workload assessment. The most important feature of the methodology is, it allows the communication requirements of mission segments to be identified by decomposing complex tasks into their component parts. The methodology represents a simple, cost-effective technique for front-end analysis of communications systems that can provide a baseline for determining the amount of information a particular communication system is capable of transmitting. We plan on applying the method in a collaborative study with Rich McKinley in the Human Systems Directorate at WPAFB, OH to determine the effects of communication jamming on team situational awareness. Another application would be to evaluate 3-D speech localization cuing to enhance situational awareness in air weapons controllers to aid in spatial information-processing. However, the validity, reliability, and sensitivity of the new methodology remain to be evaluated.

Multiple Levels of Measurement

One of the reoccurring errors of measurement methodology that was observed by Eddy (1989) during an extensive literature search of team performance measures that was either overlooked or ignored, was the failure to consider the hierarchical level of analysis for the construct being measured. For example, team coordination cannot be measured until the individual team member's performance is related to the team goal and expected outcome measure. The multilevel classification of performance measures has the advantage of placing metrics into logical subordinate and superordinate groups that indicate the predictive relationships among them.

A hierarchical framework of multiple measures of performance was developed by Clark Schingledecker as reported in Schiflett, Strome, Eddy, & Dalrymple (1990). Eddy (1990) and Dalrymple (1990) further refined the 4-tiered approach to performance measurement by adding process and outcome measures specifically related to AWACS defensive counter air mission. At the first tier are measures of *individual capability* that include single task measures of perceptual, cognitive, and motor skills which all require an active working memory. Some of the tasks are taken as a battery of performance tests external to the C^3 mission scenario sessions. Rate of responding in the form of throughput measures is commonly calculated at this level. Selected tests are embedded into the scenario and others appear as low-priority secondary tasks to measure the level of workload. An extensive database of co-variate information is usually acquired on each subject before, during, and after each session using peer rating scales, questionnaires, personality tests, and work experience.

The second class of individual-level measures focus on a single crew member's assigned role or area of responsibility. *Task-specific* measures are accumulated in real-time from individual patterns of switch actions, verbal communications, videotape recordings, and vocal stress analysis. Determinations are then made as to the extent to which the individual did or did not accomplish the specific duties as a team member with regard to target detection, identification, interception, and destruction. Over 100 measures of this type have been collected for each team member. The measures are then reduced by cluster analysis techniques and assigned weighted coherence values along the dimensions of accuracy and latency.

The third level of analysis included process and outcome measures of *system/team performance* which reflect the degree the team as a whole accomplished tasks necessary for mission success. Examples are the ratio of successful pairing of interceptors with targets and the resultant kill-ratios for each scenario segment. Measures of system performance were those measures at the team level that do not vary according to specific mission (i.e. defensive versus offensive). For example, the accuracy and speed of data transfer to interceptor pilots.

The fourth level of performance measurement is assessment of *mission effectiveness* from the Battle Area Commander's perspective. For example, if the mission is defensive counter air with protection of assets as the primary objective, then appropriate measures would include: Number of enemy infiltrations into friendly air space, amount of fuel and weapons expended, and ratio of enemy lost to friendly assets. A composite scoring scheme was developed to provide a standard quantitative measure of a team member's overall performance in relationship to the mission objectives.

This multi-level measurement system provides an implicit underlying structure that weights the significance of each measure to the others. That is, each level of the hierarchy contains groups of measures that jointly determine the measures available at the next level higher in the framework. Examining the performance measure hierarchy further, reveals that measures at each of the levels differ in their *sensitivity, generalizability,* and practical *interpretability.* It is obvious the data provided at the 4th tier (highest) is easily interpreted, while that from the lower levels offers information increasingly remote from the ultimate criterion of mission success or failure. However, this disadvantage is countered by the fact that measures at the lowest level, 1st tier, are the most sensitive and most generalizable. For example, while kill ratios are direct indices of Mission Effectiveness, these measures are influenced by a host of individual factors that make them insensitive to small but significant variations in such measures as individual decision time. Furthermore, Mission Effectiveness measures are highly specific to the individual characteristics of the test scenario. Hence, an effectiveness metric obtained under 1 set of conditions may give little indication of the system's performance in a different situation. Conversely, a measure of operator reserve capacity, such as a response time on an embedded secondary task measure, is difficult to relate directly to a criterion such as survivability. At the same time however, such a measure is generalizable across a wide range of simulation scenarios and will be extremely sensitive to variations in operator capability.

The proposed multi-level approach to performance measurement was validated in a series of complex experiments evaluating the effects of classes of antihistamine drugs on aircrew performance (Nesthus, Schiflett, Eddy & Whitmore, 1991; Eddy, Dalrymple, & Schiflett, 1992). It was found that while individual capabilities and performance can be high, and the team works as effectively as possible, the team may still fail in its mission, in conditions of high threat and high workload. The sensitivity of measures was verified since most of the degradation of the sedative type antihistamines was found at the individual level on specific cognitive tasks and specific areas of assigned responsibility and not on team or mission effectiveness measures. That is, other team members were able to compensate for the loss of capability of individual members and still succeed.

Barrett (1993) has noted in an excellent review of military research in tactical team decision-making that most of the measures of team performance developed so far have been primarily outcome measures. However, the research team within the Sustained Operations Performance Branch at Brooks AFB, Texas is now investigating individual process measures of effectiveness, to identify patterns of team interaction which lead to successful team results. The 2 general types of process measures being analyzed by Dalrymple, Eddy & Schiflett, (in press) are: (1) Task-oriented measures such as decision strategies and team workload measures, and (2) Measures related to the maintenance of team coordination through communication. This multi-level approach to individual, team/system, and mission effectiveness performance measurement allows maximum generalization to the field due to the close mapping of the embedded events in the scenario with actual wartime scenarios and tasks.

A core set of multi-level dependent measures will be integrated into the design of all experiments conducted during this basic research program to determine team performance. The dependent measures for evaluating the role of communications in team performance are presented in the next section.

Communication Measures

Verbal communications will be transcribed and treated as interactive sequences of speech events in which statements spoken by 1 team member are considered within the context of the other team members' prior and subsequent speech. The patterns of communication will be formatted into transition frequency matrices for different categories of speech as developed by Kanki & Foushee (1989); and Kanki, Lozito, and Foushee (1989). Their analytical method will be expanded to 3 or more team members rather than only a 2-sided dialogue restricted to dyad teams. Each team member's verbal interaction will be categorized by initiator and responder as follows:

<u>Initiator</u>
Demand - required action to be taken
Request - asking for some action
Question - information requests
Observation - task-related statement
Dysfluency - non-task statement or self talk
<u>Responder (solicited or unsolicited)</u>
Reply - answer to request or question
Acknowledgment - recognition of transmission
No response - not time limited

Dependent measures of verbal communications will be further developed to evaluate established conventions of speech (format, rules) and content of information received in relationship to the following:

1. Rule based compliancy to standard operating procedures
2. Clarity of transmission (intelligibility)
3. Accuracy (ratio of misinformation to correct)
4. Relevancy (only what team member needs to do their job)
5. Pacing (number of words per transmission)
6. Timeliness (optimum temporal information exchange)

Speech communication data compression techniques will be implemented using various discourse analysis software programs e.g., Petri-net diagrams, Pathfinder linkages, to detect and analyze patterns of team communications.

SUMMARY

A scientific overview of a 3-year basic research program has been outlined that will foster scientific collaborations with government and university laboratories. The research will focus on the underlying mechanisms of adaptive team coordination to gain an understanding of the role of communications in enhancing and maintaining situational awareness in distributed team decision-making. The program overview discussed the research domain, scientific goal, sub-goals, approach, research paradigm, study objectives, measurement methodology, and verbal communication dependent measures.

By working closely with a team of basic researchers from academia, government, and industry that represent diverse scientific and technical knowledge domains, the Air Force will gain a unique perspective in understanding the role of communications in improving team coordination. At the conclusion of this program, a team situational awareness database and measurement methodology will be transitioned to more advanced exploratory research programs on team performance assessment.

REFERENCES

Bales, R.F. (1950). *Interaction process analysis: Theory, research, and application.* Reading, MA: Addison-Wesley.

Barrett, L.E. (1993). *Decision-making teams: Their study in the U.S. Military.* Technical Report. *No. 93-1* (April): ONR Grant No. N00014-90-J-1786, Management & Industrial Psychology, Michigan State University, East Lansing, MI.

Broersma, T.J. (1992). *Organizational learning and aircrew performance.* Dissertation, Department of Education, University of Texas. Austin, Texas.

Dalrymple, M.A. (1990). *Development of realistic and complex command, control, and communications scenarios, and decision aids.* Proceedings of the Seventh Annual workshop on Command and Control Decision Aiding. Air Force Institute of Technology, Wright-Patterson AFB, Dayton, OH.

Dalrymple, M.A. (1991). *Evaluating AWACS strategy and tactics as they relate to simulated mission events.* Technical Paper AL-TP-1991-49, Armstrong Laboratory, Brooks AFB, Texas.

Dalrymple, M.A., Eddy, D.E., Schiflett, S.G. (In progress). *Comparative effects of antihistamines on aircrew individual performance under sustained operations.* Technical Report *AL-TR-1993-XX*, Armstrong Laboratory, Brooks AFB, Texas.

DaPolito, K.L., Jones, M.E., & Hottman, S.B. (1989). *Operator performance measures for assessing voice communication effectiveness.* AAMRL-SR-89-500, Armstrong Aerospace Medical Research Laboratory, Wright-Patterson AFB, OH.

Dyer, J.C. (1984). Team research and team training. State of the art review. In F.A. Muckler (Ed.), *Human Factors Review* (285-323). Santa Monica, CA: Human Factors Society.

Eddy, D.E. (1989). *Selected team performance measures in a C^3 environment: An annotated bibliography.* Technical Report USAFSAM-TR-87-25, USAF School of Aerospace Medicine, Brooks AFB, Texas.

Eddy, D.E. (1990). *Performance-based measures in individual and team complex decision-making.* Proceedings of the Seventh Annual Workshop on Command and Control Decision Aiding. Air Force Institute of Technology. Wright-Patterson AFB, Dayton, OH.

Eddy, D.E., Dalrymple, M.A., Schiflett, S.G. (1992). *Comparative effects of antihistamines on aircrew mission effectiveness under sustained operations.* Technical Report AL-TR-92-0018, Armstrong Laboratory, Brooks AFB, TX.

Endsley, M.R. (1988). *Design and evaluation for situation awareness enhancement.* Proceedings of the Human Factors 32nd Annual Meeting, 97-101.

Fitts, P.M. & Posner, M.I. (1967). *Human Performance.* Belmont, California: Brooks/Cole.

Foushee, H.C. (1984). Dyads and triads at 35,000 feet: Factors affecting group process and aircrew performance. *American Psychologist,* 39(8), 885-893.

Foushee, H.C. & Manos, K.L. (1981). *Information transfer within the cockpit: Problems in intracockpit communication.* In C.E. Billings and E.S. Cheaney (Eds.), Information transfer problems in aviation systems (TP 1875). Moffett Field, CA: NASA-Ames Research Center.

Foushee, H.C & Helmreich, R.L. (1989). Group interaction and Flight Crew performance. In E.L. Wiener & D.C. Nagel (Eds.) *Human Factors in Modern Aviation.*

Hottman, S.B., DaPolito, K.D., Dalrymple, M.A., McKinley, R. (In press). *Task and information analysis of selected USAF communication tasks.* Systems Research Laboratories, Inc.

Ilgen, D.R., Major, J.H., Hollenbeck, J.R. & Sego, D.J. (1991). *Decision-making in teams: Raising an individual decision-making model to the team level.* TR-91-2, Michigan State University, East Lansing, MI.

Jones, W.J. (1990). *An evaluation of voice stress analysis techniques in a simulated AWACS environment.* Masters Thesis, Industrial Engineering, Texas A&M University.

Kanki, B.G. & Foushee, H.C. (1989). Communication as a group process mediator of aircrew performance. *Aviation, Space, and Environmental Medicine*, 60, 402-410.

Kanki, B.G., Lozito, S., & Foushee, H.C. (1989). Communication indices of crew coordination. *Aviation, Space, and Environmental Medicine*, 60, 56-60.

Klein, G.A. & Thordsen, M. (1990). *A cognitive model of team decision making.* Proceedings of the 1990 Symposium on Command and Control Research, 270-274.

Klein, G.A. (1993). A recognition primed decision model of rapid decision-making. In G. Klein, J. Orasanu, R. Calderwood & C.E. Zsambok (Eds.). *Decision making in action: Models and methods.* Norwood, NJ: Ablex Publishing Corp.

Klinger, D.W., Andriole, S.J., Militello, L.G., Aldelman, L., Klein, G.A., & Gomes, M.E. (1993). *Designing for performance: A cognitive systems engineering approach to modifying an AWACS human-computer interface.* Technical Report, F33615-90-C-05332, Klein Assoc., Inc. Fairborn, OH.

Morgan, B.B., Glickman, A.S., Woodard, E.A., Blaiwes, A.S. & Salas, E. (1986). *Measurement of team behaviors in a Navy environment.* Technical Report, NTSC TR-96-014, Naval Training Systems Center, Orlando, FL.

McGrath, J.E. (1984). *Groups: Interaction and performance.* Englewood Cliffs, NJ: Prentice-Hall.

Nesthus, T.E., Schiflett, S.G., Eddy, D.R., Whitmore, J.N. (1991). *Comparative effects of antihistamines on aircrew performance of simple and complex tasks under sustained operations.* Technical Report AL-TR-91-0104. Armstrong Laboratory, Brooks AFB, TX.

Orasanu, J. & Salas, E. (1993). Team decision making in complex environments. In G. Klein, J. Orasanu, R. Calderwood, & C.E. Zsambok (Eds.) *Decision-making in action: Models and methods.* Norwood, NJ: Ablex Publishing.

Rouse, W.B., Cannon-Bowers, J.A. & Salas, E. (1992). The role of mental models in team performance in complex systems. (Manuscript submitted for publication).

Ruffell Smith, H.P. (1979). *A simulator study of the interaction of pilot workload with errors, vigilance, and decisions.* NASA TM-78482. Moffett Field, CA: NASA Ames Research Center.

Schiflett, S.G. (1990). *A triservice facility for technology transfer: A performance gateway for decision aids.* Proceedings of the Seventh Annual Workshop on Command and Control Decision Aiding. Air Force Institute of Technology. Wright-Patterson AFB, Dayton, OH.

Schiflett, S.G., Strome, D.R., Eddy, D.R., Dalrymple, M.A. (1990). *Aircrew evaluation sustained operations performance (AESOP): A triservice facility for technology transition.* Technical Paper USAFSAM-TP-90-26, USAF School of Aerospace Medicine, Brooks AFB, TX.

Shannon, C.E. and Weaver, W. (1964). *The Mathematical Theory of Communication.* Urbana, IL: University of Ill. Press (original work published 1949).

Siegel, A.I. & Federman, P.J (1973). Communications content training as an ingredient in effective team performance. *Ergonomics*, 16, 403-416.

Strome, D.R. (1990). *A unique command and control research facility.* Proceedings of the Seventh Annual Workshop on Command Decision Aiding. Air Force Institute of Technology: Wright-Patterson, AFB, Dayton, OH.

Swezey, R.W. & Salas, E. (Eds.) (1992). *Teams: Their training and performance.* Norwood, NJ: ABLEX.

Taylor, R.M. (1989). *Situational awareness, rating technique (SART): The development of a tool for aircrew systems design.* AGARD AMP Symposium on Situational Awareness in Aerospace Operations, CP478, (3/1-17).

Vincent, G.A. (1993). A new approach to command and control: The cybernetic design. *Airpower*, 7(2), 24-38.

Wickens, C.D. (1984). *Engineering Psychology and Human Performance.* Columbus, OH: Merrill Publishing.

Proceedings of International Symposium on Spoken Language, (Edited by K. Shirai), Clark, H.H., Managing problems in speaking, (1993), with kind permission from ISSD, Waseda University.

PART 1: PRESENTATIONS

MANAGING PROBLEMS IN SPEAKING

Herbert H. Clark
Department of Psychology, Stanford University

INTRODUCTION

When people talk, they manage any problems they discover quickly, skillfully, and without apparent effort. These problems arise in everything they do, from maintaining attention to maintaining face. Some result in disfluencies—pauses, repairs, fillers (like "uh" and "um"), word fragments, fresh starts—but others result in a variety of other phenomena. How are these problems managed? A common view is that speakers monitor for them and repair them when they discover them. In this paper I suggest that this view is too narrow. Managing problems is really part of a larger system in which repairs are only one strategy.

Language use is fundamentally a joint activity, and that is reflected in the way problems are managed (Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986; Schegloff, Jefferson & Sacks, 1977). When Ann and Bob converse, they each perform individual actions—e.g., uttering words, identifying sounds—but many of these actions are really parts of actions performed by the *pair* of them Ann-and-Bob. I will call actions by the pair Ann-and-Bob *joint* actions, and I will call Ann's and Bob's individual actions within them *participatory* actions (Clark and Carlson, 1982; Clark & Schaefer, 1989; Cohen, Morgan, & Pollack, 1990). In conversation—the fundamental site of language use—speaking and listening are participatory actions.

Ann's actions in talk aren't independent of Bob's, or vice versa, and that goes for their problems as well. When Ann needs extra time to plan an utterance, that isn't her problem alone. The time she needs belongs to Ann-and-Bob, so she has to coordinate with Bob on her use of that time. Likewise, when Bob doesn't understand Ann, the problem isn't his alone or hers alone. It is Ann-and-Bob's, and it takes the 2 of them working together to fix it. There are two principles here: (1) the problems that arise in language use are really *joint* problems; and (2) dealing with these problems requires *joint* management.

To complicate things, problems arise at several levels of conversation. Suppose Ann is saying something to Bob. Here are 3 levels of action, starting at the bottom:

1. *Vocalization and attention.* At the lowest level, Ann vocalizes sounds, getting Bob to attend to those vocalizations. She cannot vocalize those sounds unless she has Bob's attention, and Bob cannot register her vocalizations without attending to them. That takes Ann's and Bob's coordination.
2. *Presentation and identification.* One level up, Ann presents an utterance for Bob to identify. She must be sure Bob has identified the utterance she has presented, and he must be sure of it too, and that also takes coordination.
3. *Meaning and understanding.* One more level up, Ann gets Bob to understand what she means by her utterance. The 2 of them must reach the mutual belief, called the grounding criterion, that Bob has understood what Ann meant well enough for current purposes (Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986).

All 3 levels consist of joint actions. They each require Ann and Bob to coordinate on their individual actions. At each level, the problems Ann and Bob have as individuals are also problems for their joint action.

Problems in conversation are like infections: People prefer to deal with them before they grow into something worse. People's strategies for managing problems in conversation are much like physicians' strategies for managing infections:

1. Preventatives. These are like inoculations in averting anticipated but avoidable problems.
2. Warnings. These are like palliatives in helping participants prepare for anticipated but unavoidable problems.
3. Repairs. These are like antibiotics in remedying problems that have already appeared.

In conversation as in medicine, 1 is preferred to 2, and 2 to 3, all other things being equal. At least, this is the claim. In this brief paper, I will only allude to the evidence.

Vocalization and Attention

At level 1 (vocalization and attention), Bob must attend to Ann while she vocalizes her utterance or they will fail. Joint actions like this depend on the participants doing their parts, so for Ann and Bob to be sure of success, they need evidence that they are each doing their parts. Ann should look for evidence that Bob is attending to her, and he should try to provide that evidence. Consider this invented example:

Ann: Bob
Bob: [3 sec of no response]
Ann: Bob [louder]
Bob: What?

Ann tries to summon Bob with her first utterance, but gets no response. She takes that as evidence that Bob wasn't attending to her vocalization, a problem she has to repair. She does that by repeating the summons—only louder to capture his attention. This time he responds, giving her evidence that she has succeeded.

Whose problem is this—Ann's or Bob's? Neither of them can be held solely responsible. The problem arose from the mis-coordination of Ann's vocalization and Bob's attention. Perhaps Ann should have been more certain of Bob's attention before vocalizing, or he should have been paying closer attention, or both. In any event, Ann and Bob's joint action led to a joint problem, which required a joint remedy.

If Ann and Bob had worked together, they might have avoided the problem in the first place. There are effective preventatives for just this purpose. Several such strategies have been described by Goodwin (1981), one of which is illustrated here with Lee talking to Ray:

Lee: Can you bring- (0.2) Can you bring me here that nylon?

As a videotape of this utterance shows, just when Lee wants to start speaking, he sees that Ray is looking away. If Lee were to start his utterance, Ray wouldn't be attending, and that would create a problem they would later have to repair. Lee's strategy is to prevent the problem by using "can you bring" to request Ray's attention and by starting again only once he has Ray's attention. Indeed, Lee restarts "can you bring" precisely as Ray begins to turn his head toward Lee. Lee's strategy, which itself requires a joint action, was designed not to remedy an existing problem, but to prevent a future problem from arising.

Presentation and Identification

At level 2 (presentation and identification), Ann must present an utterance, getting Bob to identify it, and that again requires coordination. For them to succeed, Ann needs evidence that Bob is identifying her utterance, and he needs to provide that evidence. The evidence addressees provide may show they haven't yet identified an utterance at all, a problem speakers usually repair by repeating the utterance, as in this spontaneous example (from Svartvik & Quirk,1980):

A: ((where *are* you))
B: m? .
A: where are you
B: well I'm still at college
A: [continues]

Or the evidence may show that the addressees have identified only part of an utterance, as here (from Svartvik & Quirk,1980):

Roger: now, - um do you and your husband have a j- car?
Nina: - have a car?
Roger: yeah
Nina: no -

Or the evidence may show that the addressees have misidentified all or part of a presentation, a problem speakers usually correct by repeating the misidentified part, as in this exchange of an address (from Svartvik & Quirk,1980):

A: yes forty-nine Skipton Place
B: forty-one
A: nine . nine
B: forty-nine, Skipton Place,

In all 3 examples, the problems are joint ones, and they are managed with joint remedies.

Speakers may discover problems from evidence provided by addressees, as in these examples, but also from monitoring their own presentation, as in this example (from Svartvik & Quirk,1980):

Ann: they still talk about rubbish tins, which is the American the Australian
Beth: yeah
Ann: expression, . for that thing you put all the . stuff in at the back gate, you know

Ann catches the error in "American" on her own and instantly repairs it to "Australian." Immediate self-corrections like this are preferred for at least two reasons. First, they aren't as costly—they require only an extra word or phrase instead of two extra turns. And second, although they repair one problem, they prevent deeper and more costly misunderstandings down the line. They are not only repairs, but also preventatives.

Speakers anticipate some problems even before they are manifest. For example, speakers recognize that most presentations have an *ideal delivery*—one that is fluent, correct, and optimal for identification (Clark & Clark, 1977). They also recognize that any deviation from the ideal may cause their addressees problems, so they should try to achieve the ideal delivery. The trouble is, they usually cannot formulate an entire presentation before they begin speaking. They are forced to formulate one phrase at a time, interrupting their utterances to do that. Since they recognize that interruptions and pauses pose problems for their addressees, how should they proceed?

If speakers foresee a delay or interruption even when they cannot prevent it, they can help their addressees prepare for it by warning them about it. One way is by signaling the onset of an interruption. Surprisingly, they can also signal its size. Evidence shows that speakers use "uh" to signal short interruptions, and "um" to signal more serious ones. When 25 university students were asked 40 questions like "What is the name of the first man to run a mile in under four minutes?" In conversational settings, there was often a delay in their answers. If they began without a filler,

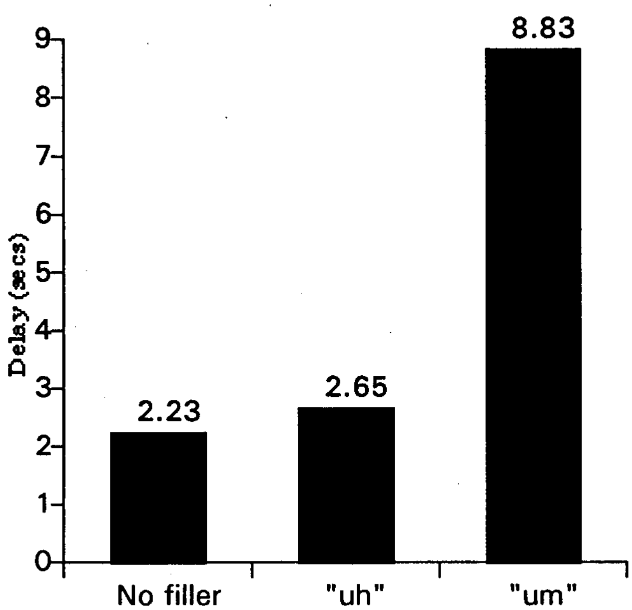


FIGURE 1: Answer delays as a function of filler use

the delay averaged 2.23 seconds; if they began with "uh," it averaged 2.65 seconds; but if they began with "um," it averaged 8.83 seconds (Smith & Clark, 1993). The delays of answers are shown in Figure 1.

In a study of the London-Lund corpus of English conversation (Svartvik & Quirk, 1980), Fox Tree and I computed the percentage of times that "uh" and "um" were preceded and followed by perceptible pauses. The percentages are summarized in Figure 2.

Speakers quite often produced "uh" and "um" after pauses. But they tended to use "uh" when there were no further pauses and "um" when there were. So in both studies, speakers used "uh" and "um" to warn addressees about the size of interruption they were anticipating.

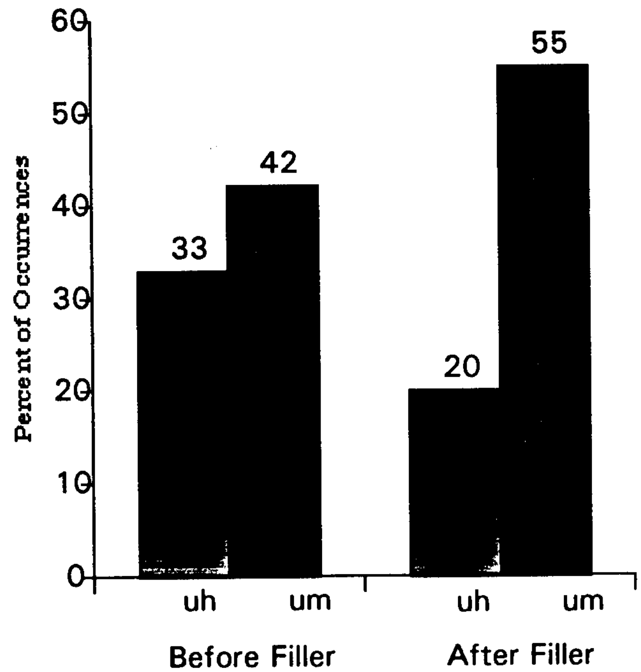


FIGURE 2: Occurrence of pauses before or after fillers

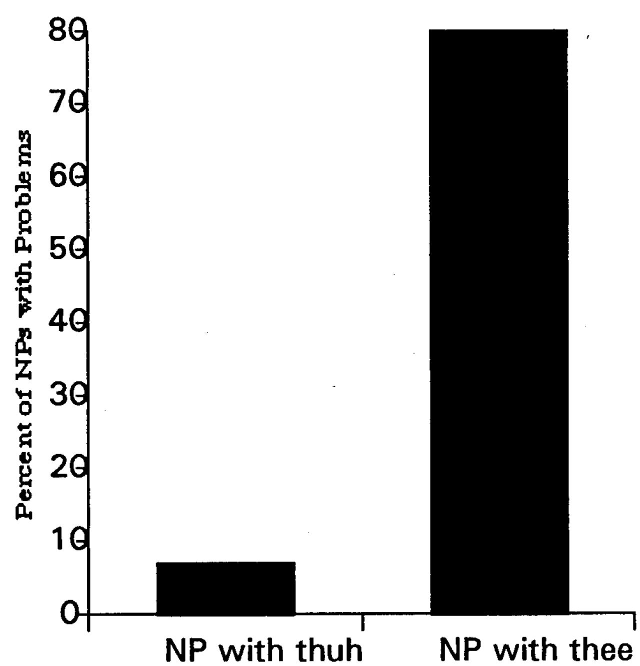


FIGURE 3: "Thuh" or "Thee" preceding noun phrases with problems

Speakers also warn addressees about problems in formulating noun phrases (NPs). Although "the" is ordinarily pronounced "thuh," it is sometimes pronounced "thee" when speakers foresee a problem in formulating the current NP. In the London- Lund corpus, Fox Tree and I found disruptions in 7% of the NPs introduced by "thuh," but in 80% of those introduced by "thee." The percentages are shown in Figure 3.

Apparently, speakers choose "thee" to warn of an approaching disruption, and that should help addressees prepare for it.

At the level of presentation and identification, then, the participants Ann and Bob not only repair existing problems, but try to prevent future problems and warn of approaching but unavoidable problems.

Meaning And Understanding

At level 3 (meaning and understanding), Ann must get Bob to understand what she means with her utterance. To succeed, they must reach the mutual belief that he has understood her well enough for current purposes, and for that, he must provide her with evidence of his understanding. When speakers detect misunderstandings in the evidence provided by addressees, they initiate the needed repairs, as here (from Svartvik & Quirk, 1980):

B: k who evaluates the property - - -
A: uh whoever you asked, the surveyor for the building society
B: no, I meant who decides what price it'll go on the market -
A: (- snorts) whatever people will pay - -

When A's answer to B's question shows that A has misunderstood him, B starts his correction "No, I meant." In other cases, addressees detect the problems first and ask for repairs, as B does here (from Svartvik & Quirk, 1980):

A: Well wo uh what shall we do about uh *this* boy then
B: Duveen?
A: m
B: well I propose to write, uh saying . I'm very sorry [continues]

When B isn't certain which boy A is referring to, he gets A to confirm that it is Duveen.

It is even more prevalent for speakers to find problems in their utterances and repair them before they cause further misunderstanding, as in this example (from Svartvik & Quirk, 1980):

Jane: this is the funny thing about academics, - that if you're no- uh you know, I I've come to it, so late, I mean I've had a lifetime of experience, rolling around,

At one point, Jane says "that if you're no-" then cuts herself off and then offers the repair "I I've . come to it, so late." She then offers a second repair, "I've had a lifetime of experience, rolling around." But she does more than make the repairs. She signals the type of repairs they are by means of *editing terms*—"uh you know" for the first and "I mean" for the second. This way she helps her partner prepare for the repairs with warnings about when they are coming and why.

Speakers have other less obvious strategies for preventing misunderstandings. One is the use of hedges such as "kind of," "sort of," and "like." Consider "sort of" in this example (from Svartvik & Quirk, 1980):

Reynard: but you see, it is sui generis, so ((it'll)) so . anybody who is looking for, um . a a niche to fit it a ready-made niche, . in English grammar to fit it . into, . is sort of begging for the moon, - - ((you see))

If Reynard had said "is begging for the moon," he would have implied that the expression captured precisely what he meant. By saying "is sort of begging for the moon," he implies that the expression is only approximately right. Evidence shows that when speakers use hedges such as "kind of," "sort of," and "like," they are indeed less accurate. When university students were asked to retell stories they had just heard, they produced direct quotations both with and without hedges (see Figure 4). When they

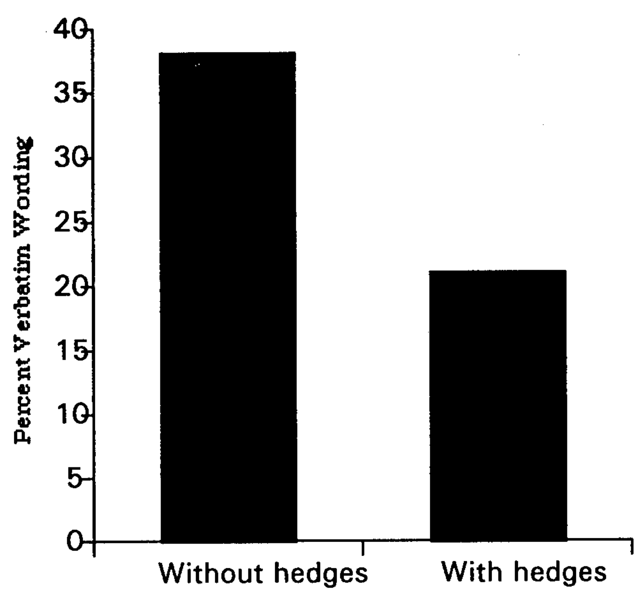


FIGURE 4: Verbatim wording with and without hedges

didn't use hedges, they reproduced 38% of the verbatim wording from the original stories. When they did use hedges, the percentage was only 21% (Wade & Clark, 1993).

These speakers were right to warn their addressees of their inaccuracies.

CONCLUSION

In a common view of language use, problems are managed by speakers, who monitor for them and repair them when they arise. I have argued that this view is too narrow. For one thing, managing problems is something the participants do together. All problems are ultimately joint problems, and they have to be managed with joint strategies. For another thing, speakers do more than make repairs. They have strategies for *preventing* certain problems from arising at all. For problems that are unavoidable, they have strategies for *warning* their partners—to help them prepare for the problems. And for problems that arise anyway, they work with their partners in repairing them. In the management of problems, preventatives are preferred to warnings. Repairs are the last resort.

REFERENCES

Clark, H. H., & Carlson, T. B. (1982a). Hearers and speech acts. *Language*, 58, 332-373.

Clark, H. H., & Clark, E. V. (1977). *Psychology and Language*. New York: Harcourt Brace Jovanovich.

Clark, H.H., & Schaefer, E.F. (1989). Contributing to discourse. *Cognitive Science*, 13, 259-294.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1-39.

Cohen, P. R., Morgan, J. L., & Pollack, M. E. (Eds.) (1990). *Intentions in communication*. Cambridge: MIT Press.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.

Schegloff, E.A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53, 361-382.

Smith, V.L., & Clark, H.H. (1993). On the course of answering questions. *Journal of Memory and Language*, 32, 25-38.

Svartvik, J. & Quirk, R. (1980). *A corpus of English conversation*. Lund, Sweden: Gleerup.

Wade, E., & Clark, H.H. (1993) Reproduction and demonstration in quotations. *Journal of Memory and Language*, 32.

PILOT AND CONTROLLER COMMUNICATION ISSUES

Linda Connell

NASA Ames Research Center, Aviation Safety Reporting System

INTRODUCTION

The NASA Aviation Safety Reporting System (ASRS) was established in 1976. Since that time, ASRS has received, processed, and analyzed approximately 280,000 voluntarily submitted aviation safety reports from pilots, air traffic controllers, and other participants within the National Airspace System. Currently, the system is averaging 30,000 reports per year (ASRS, 1994a). The establishment of ASRS was largely influenced by the crash of TWA 514 near Washington Dulles Airport in 1974. During the course of the National Transportation Safety Board (NTSB) investigation of this unfortunate accident, it was learned that pertinent information concerning a known hazard had not reached the pilots of this fateful flight and would have likely prevented the accident. As a result, it was recommended that a program be established to provide a central, national resource for information on aviation incidents. As a means of preventing accidents, the ASRS uses the information it receives to remedy reported hazards, to conduct research on pressing safety problems, and to otherwise further aviation safety.

The ASRS provides confidentiality to pilots, air traffic controllers, and others who discuss the circumstances surrounding the occurrence of an actual aviation incident (e.g., an unsafe flight condition, an inadvertent violation of a Federal Air Regulation, a near-accident, etc.). This assurance of confidentiality is provided by NASA, an independent government agency, under the conditions and requirements established in the FAA Advisory Circular (AC No. 00-46C). In this Advisory Circular, the FAA extends limited immunity to individuals who report unintentional rule violations. This provision has been crucial to reporter motivation and confidence to report incidents in a non-threatening format. As stated in the Advisory Circular, "The filing of a report with NASA concerning an incident or occurrence involving a violation of the Act of the Federal Aviation Regulations is considered by the FAA to be indicative of a constructive attitude. Such an attitude will tend to prevent future violations." (pp. 3).

When an incident occurs, the reporter submits an ASRS reporting form which provides a detailed summary of the conditions and situation variables involved in the incident. The form includes information about the type of operation, type of aircraft, qualifications of the reporter, weather, airspace, etc. The most vivid detail of the incident event, however, is provided in the narrative section of the report where the reporter recounts the actual events preceding, during, and following the incident. This combination of information is the single, largest advantage of incident reporting to the on-going efforts of accident prevention. The reporters involved in the event are able to relate the conditions surrounding the incident, but they are also able to relate how they detected and resolved the problem in a satisfactory manner. Often accident investigations are unable to recreate this kind of information. Incident analysis can and does provide this useful information for use in targeting potential areas for improvement and thus, contribute to accident prevention.

Because of the richness of the data provided to the ASRS, the opportunity to use the information to accomplish 2 purposes is available, 1 short-term and 1 long-term. The first purpose is to identify deficiencies and discrepancies in the current aviation system (ASRS, 1994b). The data is used to inform the participants in the system of safety concerns or developing problems within the system. To satisfy the goals of this purpose, there is an alerting function in place within the ASRS which distributes de-identified information on any significant safety item to all responsible agencies and participants of the airspace system. The long-term purpose is to provide data for planning and improvements to the National Airspace System (ASRS, 1994b). ASRS maintains an active database of aviation incident reports to provide relevant information for guiding aviation human factors research efforts and recommendations for future aviation procedures, operations, facilities, and equipment. ASRS particular concern is the quality of human performance in the aviation system.

To maximally support these purposes, each incident report is reviewed and analyzed by a team of experienced aviation safety analysts. This team is composed of retired pilots and air traffic controllers from all types of operations and environments, such as commercial Part 121, commuter Part 135, corporate and general aviation Part 91, Air Traffic Control (ATC) Tower, Terminal Radar Control (TRACON), Air Route Traffic Control Center (ARTCC), and Flight Standards District Office (FSDO) organizations. The incident reports are evaluated by the analysts, selections are made for full and abbreviated processing, telephone callbacks to the reporters for clarification may be made, and each report is categorized into a selection of categories describing the incident event characteristics. One of the many areas that are evaluated in these reports is communication or information transfer, which is the focus of this paper.

INFORMATION TRANSFER

The structure and content of the current pilot/air traffic controller communication interaction is the result of an evolutionary process developed to handle the demands of the necessary aspects of information transfer within the aviation system. Whether this interaction has evolved toward the most efficient and accurate method available is often questioned when investigating communication errors. However, the established method has been successful overall and responsive to numerous aspects within the dynamic nature of the National Airspace System.

In the aviation environment, radio communication is essentially the only means of information method utilized for this communication interaction transfer between the aircraft and the ground. There is a three part transaction beginning with the initial transmission of information, usually from the air traffic controller to the pilot (Figure 1). The pilot response is called a *readback,* which includes the necessary components of the aircraft callsign and a repeat or acknowledgment of the information received (as understood by the pilot). The last part of this communication transaction is a *hearback*, which requires the air traffic controller to evaluate the pilot readback for accuracy and to clarify any discrepancies. Although new technology (e.g., automated pre-departure clearance and future data-link systems) will be introduced as an aid to information transfer, the necessary components of this transaction process will continue to be required to maintain the orderly and timely flow of information between the flight deck and the air traffic control facility.

In the initial 5 years of the ASRS program's existence, over 70% of the reports submitted noted problems in the transfer of information in the aviation system. Information transfer issues continue to represent the largest category of problems contained in these reports. Additional research by the ASRS staff into the events characterized by information transfer issues has yielded other reports and publications focusing attention on pilot/controller communications. The existence of this very pervasive issue and a discussion of the characteristics has been documented in ASRS publications on specific problems of information transfer (Billings & Cheaney, 1981), call sign confusion (Monan, 1983) and readback/hearback errors (Monan, 1986).

ASRS DATABASE OVERVIEW

The ASRS database for 1993 was searched for all incident reports categorized for communication/information transfer issues. The total database included 24,376 and the total full-form incidents included 6,844 reports (ASRS, 1994a). The information necessary to evaluate communication issues in detail is included in the analyses of the full-form reports. Therefore, the following information will focus on those reports only. Eleven communication problem areas were identified and are presented in Table 1.

The 2 most frequently reported communication problems are controller and pilot communication technique, respectively 50% and 46%. These categories are expected to be high as the classification scheme is not mutually exclusive; however, this does indicate that the evaluation of these reports describes communication technique as a large general problem area. Controller communication technique is coded when ASRS analysis indicates that a controller may have used less-than-optimum means for communicating the message. There may be a phraseology problem involved, or there may be an issue of just what information was communicated by the controller and when information was communicated (ASRS, 1994a). Pilot communication technique refers to a wide variety of communication problems fostered by pilots. Two types of pilot communication technique problems are pilot failures to monitor frequencies and pilot failures to verify doubtful communications. Many communication problems reported to ASRS arise because pilots are not "guarding" frequencies carefully. They miss clearances directed to them, or intercept clearances intended for other aircraft. Also, pilots will often admit in the ASRS reports that they had some doubt about an ATC communication, but chose to clarify it with another crewmember rather than ATC. If the other pilot was not listening carefully, the crew may end up doing what they "thought they were told," or "expected to be told." This problem seems to be rooted in frequency congestion, which makes it difficult to verify ATC communications, and the personal pride of the flight crews, which make them reluctant to admit to a monitoring failure (ASRS, 1994a).

Within the air traffic control environment, intra- and interfacility coordination are described as a contributor to communication problems in 12% of these incidents. Controller reports often cite distractions, such as inter/intrafacility coordination activities, as the reasons for not monitoring frequencies with full attention. As a result, they miss pilot call-ups or fail to detect erroneous readbacks. Controller monitoring failures can be related to workload. During busy periods, controllers may shift their attention to aircraft "B" as soon as they

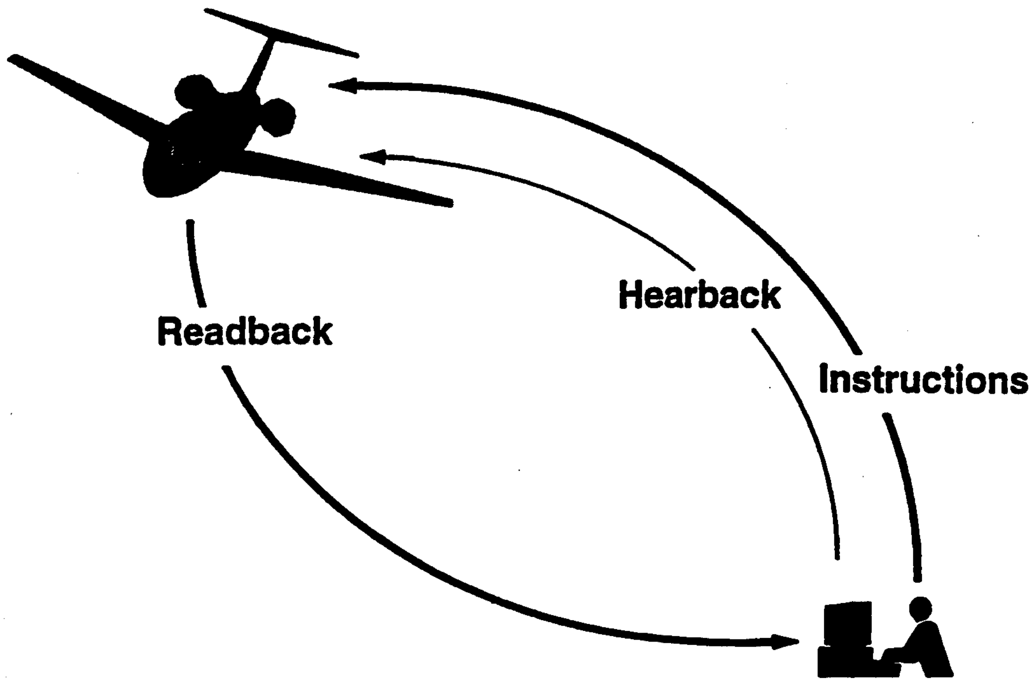


FIGURE 1: Pilot-Controller Communication

Communication Problem Areas*	Number of Citations	Percentage of Total Communication Reports
Controller Communication Technique	858	50%
Pilot Communication Technique	794	46%
Readback/Hearback	206	12%
Frequency Congestion	160	9%
Interfacility Coordination	124	7%
Phraseology	96	6%
Language Problems	92	5%
Intrafacility Coordination	87	5%
Headset/Speaker Malfunction	56	3%
Simultaneous Transmission	47	3%
Similar Sounding Alphanumerics	33	2%

***Communication problem area classifications are not mutually exclusive.**

TABLE 1: ASRS Incidents With Communication Problems (1993)

have given a clearance to aircraft "A" without waiting for a readback. The "hearback" portion of the communication transaction remains a significant and difficult task to accomplish (ASRS, 1994a).

Readback/hearback, frequency congestion, and phraseology problems are described as a contributor in these communication incidents, respectively 12%, 9%, and 6% of the time. Frequency congestion results in both miscommunications and non-communications. It encourages communication short-cuts and deviations from standard phraseology (ASRS, 1994a). It interacts with and contributes negatively to many of the other communication problem areas. In this paper, the communication problem areas of Readback/Hearback and Phraseology were further evaluated for: (1) the types of reported incident anomalies associated with these problems, (2) the evaluation of who was the original reporter, (3) who attributed to the primary problem, and (4) the phase of flight where the incident occurred.

Readback/Hearback

When an error in the readback/hearback process between the pilot and air traffic controller occurred as a contributor to a reported incident, the report was coded into this communication problem area. ASRS has a steady flow of reports that reference problems with information verification. Often, these references appear in the form of complaints by flight crews that a controller failed to correct an inaccuracy in their understanding of a clearance. On other occasions, pilots are accused of failing to readback a clearance or failing to provide a complete readback to confirm its content. The article by Morrow, et. al. (1994), investigates the complexity of the collaborative strategies used by pilots and air traffic controllers in the communication process. The hierarchy of reported incident anomalies is presented in Figure 2.

Altitude deviation, airborne conflict, less than standard separation, track or heading deviation, and runway transgression comprise the top 5 incidents that involved a readback/hearback problem as a contributor. The frequency of pilot and air traffic controller report submission is indicated in Figure 3. The person or variable determined by the report analyses to have contributed to the primary readback/hearback problem is presented in Figure 4. A flight crew and an air traffic controller could potentially contribute equally to a reported incident; however, the flight crew in the overall data is attributed with 67% of the primary problem and air traffic control with 23%. Of course, this result could be affected by the frequency of incident reporting in this readback/hearback category for pilots at 81% and air traffic controllers at 19%.

The phase of flight most often involved in the readback/hearback incident is presented in Figure 5. The top 5 phases of flight that were reported as including a readback/hearback problem are cruise, climb, descent, takeoff, and taxi. The flight phases listed below the first 5 are also descriptive of other points in the conduct of a flight where readback/hearback problems can play a role in an incident. The results in this section, as in many other variables, are probably influenced by the frequency of opportunity. In other words, the major portion of a flight is in cruise, therefore providing an increased probability of a readback/hearback error. Further in-depth analyses of each reporter's narrative could illuminate the potential factors explaining some of these findings.

Phraseology

When communication incidents are evaluated in relation to phraseology as a communication problem area, it is discovered that phraseology problems occur in virtually all types of events where instructions from ATC are involved. This category of communication problems most often refers to the deviation from standard phraseology by pilots or air traffic controllers (ASRS, 1994a). This deviation can be in message content or in delivery technique, as explained in detail in the article by Prinzo & Britton (1994).

Altitude deviation, runway transgression, and airborne conflict are consistently the 3 most frequently reported occurrences related to phraseology problems (Figure 6). Track or heading deviation is the fourth most frequent with less than standard separation, ground conflict, aircraft equipment problem, and near midair collision sharing the fifth most frequently reported type of incident related to phraseology problems.

The frequency of pilot and air traffic controller report submission is indicated in Figure 7. The person or variable determined by the report analyses to have contributed to the primary readback/hearback problem is presented in Figure 8. As in readback/hearback incidents, the flight crew reports more frequently and is attributed with the primary problem more frequently. As previously mentioned, the primary problem result could be influenced by the volume of reporting by flight crew and the apparent willingness to accept responsibility for the ensuing problem. Although, both pilots and air traffic controllers use nonstandard phraseology, pilots indicate lower levels of awareness of proper phraseology and weak adherence to communication protocols. The system provides very little formal reinforcement to the pilots to communicate with

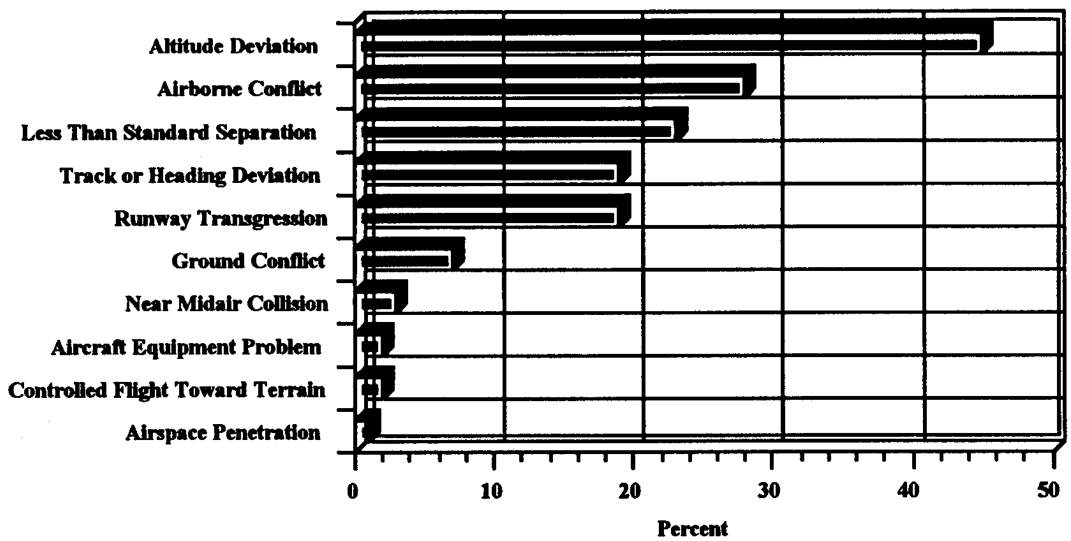


FIGURE 2: Anomalies Associated With Readback/Hearback Incidents.

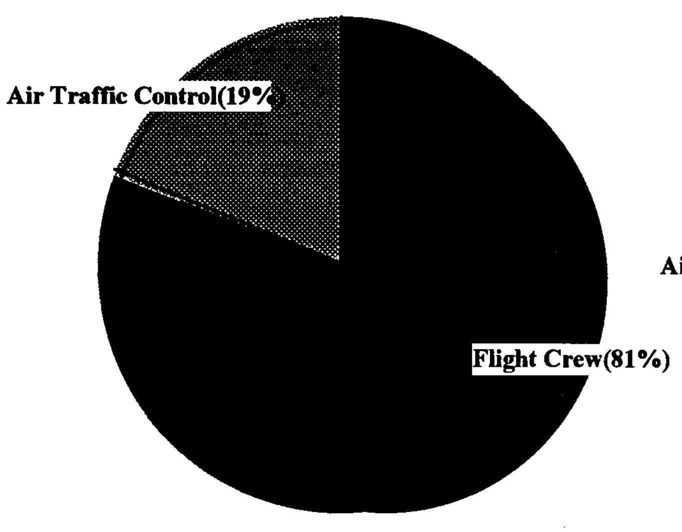


FIGURE 3: Readback/Hearback Incident Reporters.

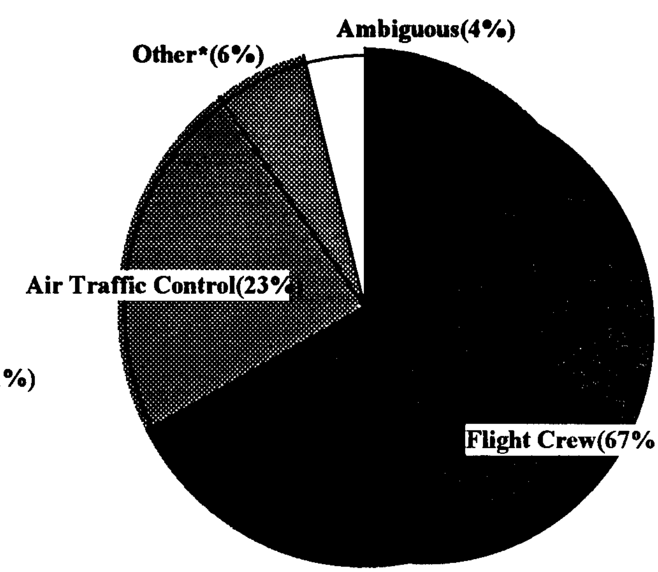


FIGURE 4: Primary Problem: Identification in Readback/Hearback Incidents.

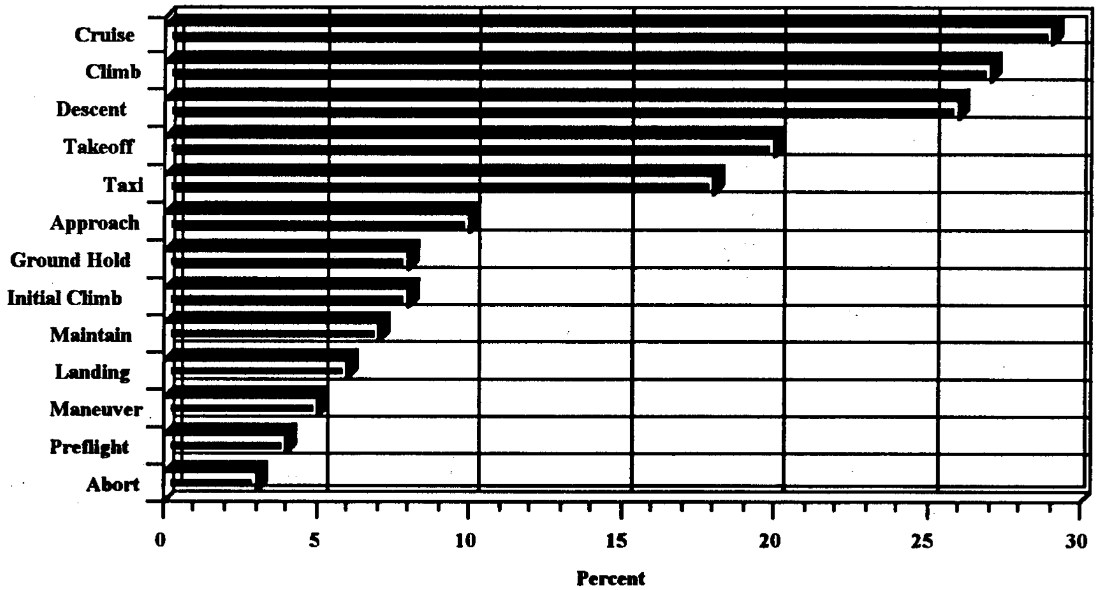


FIGURE 5: Phases of Flight Involved in Readback/Hearback Incidents

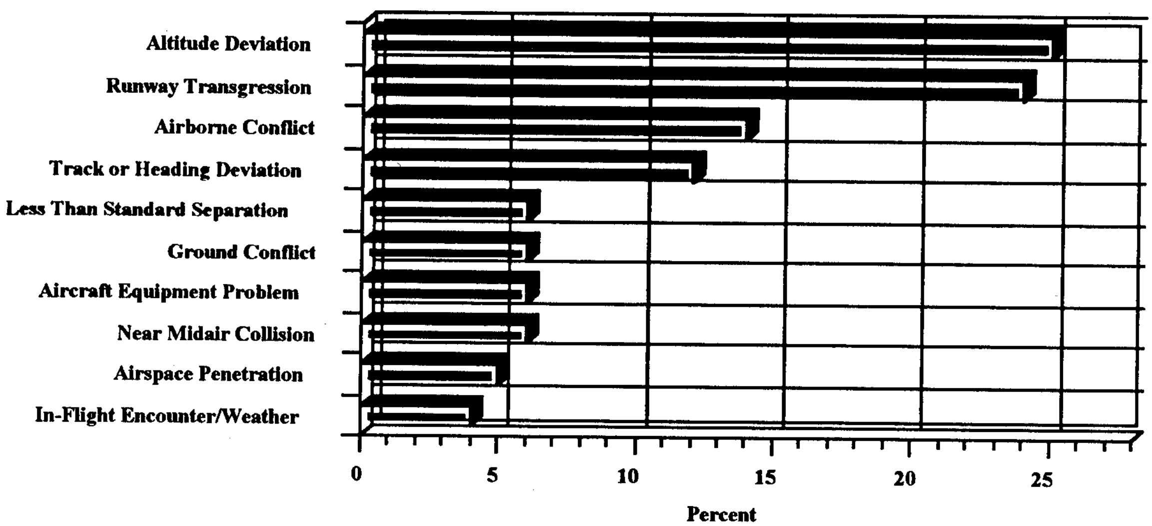


FIGURE 6: Anomalies Associated With Phraseology Incidents

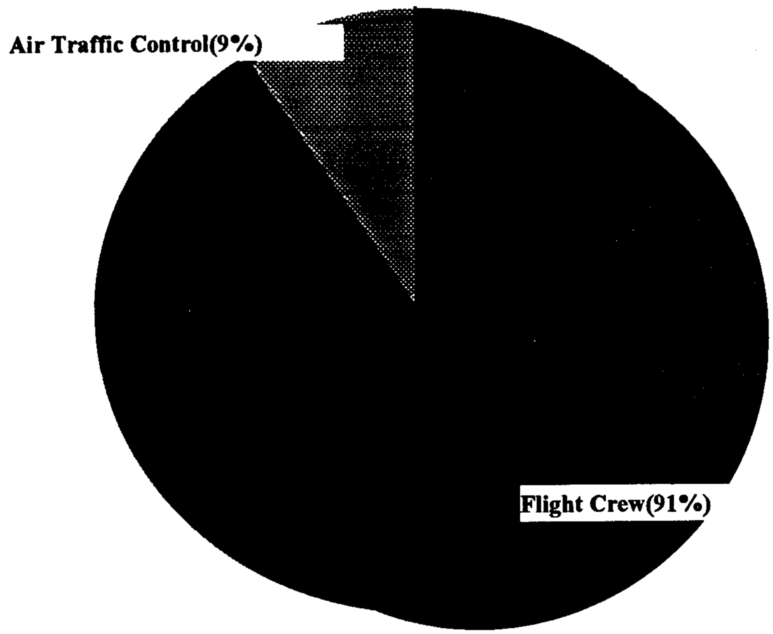


FIGURE 7: Phraseology Incident Reporters

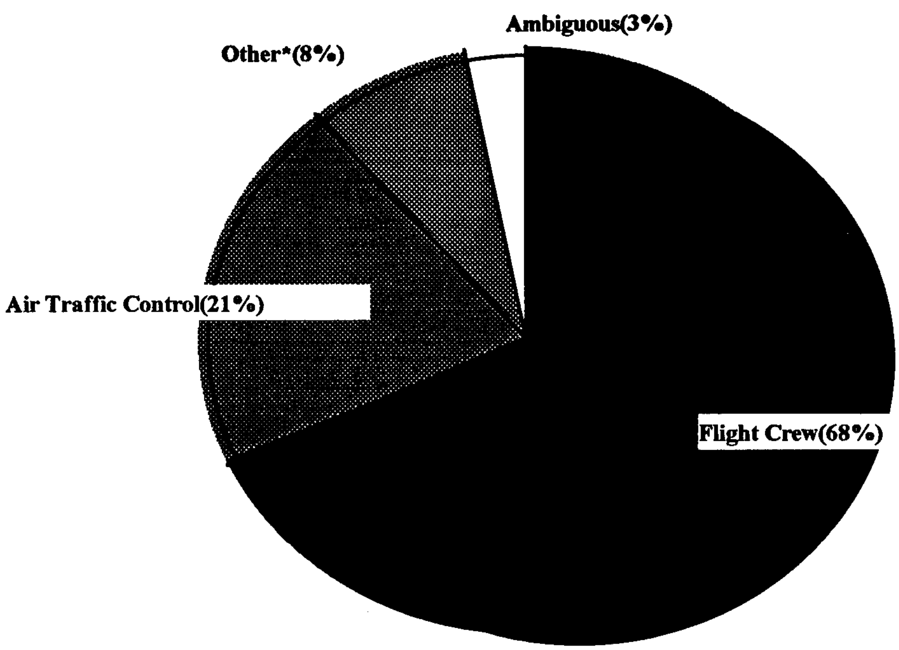


FIGURE 8: Primary Problem Identification in Phraseology Incidents.

standard phraseology. There is considerably more stringent surveillance in the ATC environment for the use of standard phraseology.

The flight phases of taxi, approach, and takeoff occurred equally in this set of phraseology incidents, followed by climb, descent, and ground hold (Figure 9). The results in this section may be identifying the crucial phases where the adherence to standard phraseology is potentially more important. A continued in-depth analyses of each reporter's narrative may provide the explanation for these findings.

SUMMARY

Communication failures among pilots and air traffic controllers are a common theme in the Aviation Safety Reporting System. This was true in 1976 when the ASRS program was instituted, and it is equally true now (Connell, 1994). The data presented in this paper reflect the reports currently available in the incident reporting database and can provide a basis for investigating these problem areas. As further confirmation of the identified communication problem areas, a recent ASRS effort concerning communication issues was conducted. A collective team of expert ASRS analysts addressed 11 pilot and air traffic controller communication issues based on their processing of thousands of communications-related incidents (Chappell, 1994). Their summary included these issues:

- Tendency of either pilots or controllers to hear what they want to hear during a readback.
- Reluctance of flight crews to question a controller or seek clarification of a clearance, especially when the controller sounded rushed, angry, or overloaded.
- Flight crew acceptance of an uncommon clearance without questioning it with ATC (e.g., climb to 13,000 off of ORD instead of the usual 5,000 ft. restriction. This was an ATC error not caught by the flight crew).
- The conveyance of a cavalier attitude by a pilot with the response, "Yeah, we'll do all that," and subsequently surprises ATC when the flight is unable to ". . . do all that."
- "Rapid Fire" transmissions from ATC often with abbreviated callsigns.
- Failure of ATC to alert or give emphasis to other flights to the similarity of callsigns.
- Flight crew failure to receive the "Golden Words" from ATC, but were under the impression that they were:

  "Cleared for takeoff"

  "Cleared to land"
- The inevitable traps caused by a lengthy ATC clearance with too many numbers.
- Issuance of taxi instructions during application of reverse thrust on landing rollout.
- Nonessential company/auxiliary communications during approach or initial departure phases.
- The use of nonstandard phraseology by both pilots and controllers:

  "Do the best you can."

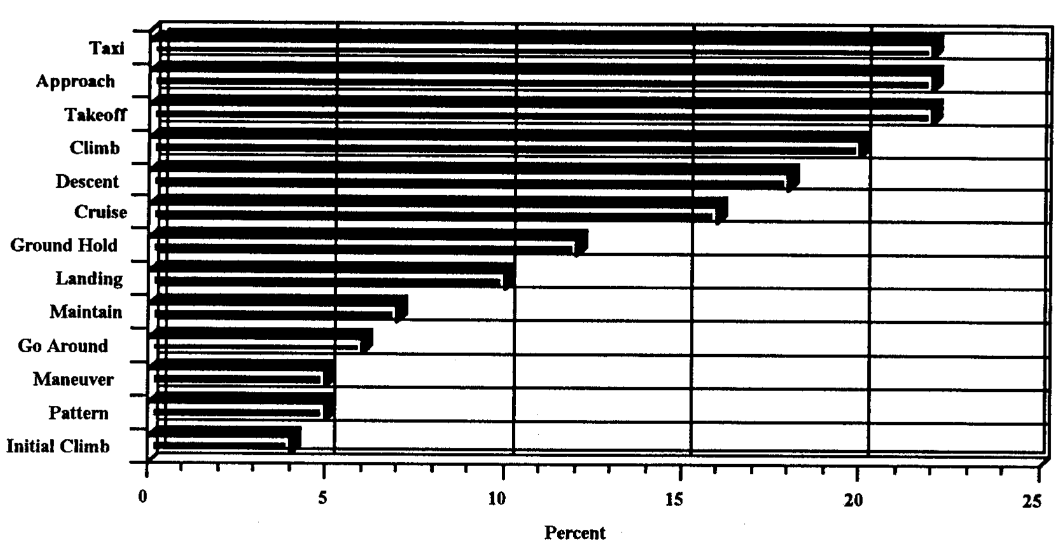
  "Give me your best rate of climb."

  "Maintain two-three-zero." (Is that speed, heading, or altitude???)

  Use of abbreviated callsigns

The database summary and the previous list point out some of the areas of concern for both pilots and air traffic controllers and is offered constructively without any attachment of blame. The overall goal is

FIGURE 9: Phases of Flight Involved in Phraseology Incidents.



collaborative, cooperative interactions between pilot and air traffic controller which underlie the equal responsibility for efficient, accurate information transfer. The system needs to support this relationship in its procedures, rules, requirements, and training. Routine training of pilot/air traffic controller techniques will reinforce effective radio skills and communication awareness. There is a need to emphasize that there is no penalty for verification, confirmation, or clarification of information whenever there is any doubt or question. There may be a penalty or worse for accepting incorrect information. Full readback/hearbacks will increase the likelihood of error detection and may reduce workload and frequency congestion overall by omitting the additional communication necessary when an error is discovered. In general, listening on the part of the pilot/air traffic controller team is an important skill needing constant reinforcement and improvement.

REFERENCES

Synopsis of Pilot/Controller Communication Reports. ASRS Reference, June 13, 1994.

Aviation Safety Reporting System-Program Summary. ASRS Reference, August 22, 1994.

Billings, C., & Cheaney, E. (1981). *Information transfer problems in the aviation system.* NASA Technical Paper # 1875.

Chappell, S. (1994). *Aviation safety reporting system - Communication in the skies.* ALPA, ERAU, NATCA, & USAir, 13th Annual Communicating for Safety Seminar, Orlando, Florida, May 19, 1994.

Monan, W. (1983). *Addressee errors in ATC communications: The call sign problem.* NASA Contractor Report # 166462.

Monan, W. (1986). *The hearback problem.* NASA Contractor Report # 177398.

Morrow, D., Rodvold, M., McGann, A., & Mackintosh, M. (1994). *Collaborative strategies in air-ground communication.* Invited Panel Presentation on Operational Voice Communication, Published Proceedings from SAE Aerotech '94.

Prinzo, V., & Britton, T. (1994). *A taxonomic analysis of terminal air traffic control/pilot communications.* Invited Panel Presentation on Operational Voice Communication, Published Proceedings from SAE Aerotech '94.

HOW TO SAY IT AND HOW MUCH: THE EFFECT OF FORMAT AND COMPLEXITY ON PILOT RECALL OF AIR TRAFFIC CONTROL CLEARANCES

Judith Bürki-Cohen
USDOT-RSPA-Volpe National Transportation Systems Center

INTRODUCTION

I am presenting preliminary results of a part-task simulation study investigating the best way for air traffic controllers to communicate numerical air traffic control (ATC) information, such as heading, radio frequency, air speed, altimeter setting, and altitude. This work was done in collaboration with Dr. Kim Cardosi at the Volpe National Transportation Systems Center and the MIT Flight Transportation Laboratory. The work was sponsored by the Federal Aviation Administration's Research and Development Service.

The study attempted to answer 2 questions:

1) Which presentation *format* best helps pilots *remember* numerical ATC information correctly?
2) How *much* information should a single transmission contain?

The first question was motivated by a recent change in ATC communication procedures. Currently, controllers are required to convey all numbers in sequential format, that is, digit by digit. For example, a speed of 310 knots has to be conveyed as "Increase speed to three one zero." Originally, it was assumed that this format is more intelligible in a noisy cockpit than the corresponding grouped format, i.e., "three hundred and ten." Recently, however, controllers have been allowed to restate altitude clearances in grouped format, e.g., "seventeen thousand niner hundred," after having given them in sequential form, i.e., "one seven thousand niner hundred." This change was based on controllers' intuitions that numbers in grouped format might be better remembered. No objective data were available, however, to motivate this change in procedure.

The second question was motivated by an analysis of enroute controller-pilot voice communications by Cardosi (1993). In this analysis of audiotapes recorded at Air Route Traffic Control Centers, she found that the more information a clearance contained, the more likely it lead to communication problems, that is, a request for repetition or an incomplete or erroneous readback of the information.

METHOD

We presented airline pilots with taped air traffic control clearances corresponding to a low-sector en route environment. They were spoken by an ATC specialist. To study the effect of complexity on pilots' recall, the clearances contained either 3, 4, or 5 pieces of information. To study the effect of format, the numbers contained in the clearance were said either in sequential, grouped, or restated format.

The pilots were asked to assume the role of the non-flying, communicating pilot. They listened to the clearances over headsets and read the clearances back into a microphone. They also set the values on a mock-up mode control panel (Figure 1). They were asked to respond in all cases, but could press a "Say Again" button next to each setting to indicate that they would have asked for a repeat in real life. They could do the readback and settings in any order. They were not permitted to use a notepad to aid memory.

The controlled laboratory setting allowed us to increase the validity of our results. The clearances were spoken very clearly and not too fast, to avoid confounding the effect of our variables on pilot recall with intelligibility. Also, to avoid contamination of our data with pilots' expectations, the clearances did not follow a realistic flight scenario. We did, however, use the terms "reduce," "increase," "descend," and "climb" appropriately. Also, we restricted the clearances to possible values only and observed speed/altitude restrictions and rules such as pairing even altitudes with western headings and odd altitudes with eastern headings.

Another factor that might contaminate the results is the context and order in which the information is presented. Observing the constraints that altimeter readings follow altitude changes and frequency is given last, all possible combinations and orders of information were carefully counterbalanced across the 3 formats and complexity levels. This resulted in 36 different clearances that pilots had to respond to, 12 at each complexity level. Each experimental clearance was complemented by 2 similar (with respect to order and complexity) clearances for other aircraft, resulting in a total of 108 clearances (actually, 110 clearances, counting the 2 "catch" clearances for the pilot containing unexpected information such as traffic point-outs, etc.). To present each clearance in all 3 presentation formats, pilots were tested on this set of experimental and filler trials 3 times. The sessions differed in the order of presentation of the clearances, with the constraint that no 2 clearances for the pilot immediately followed each other. Also, constraints such as the even altitude/western heading had to be

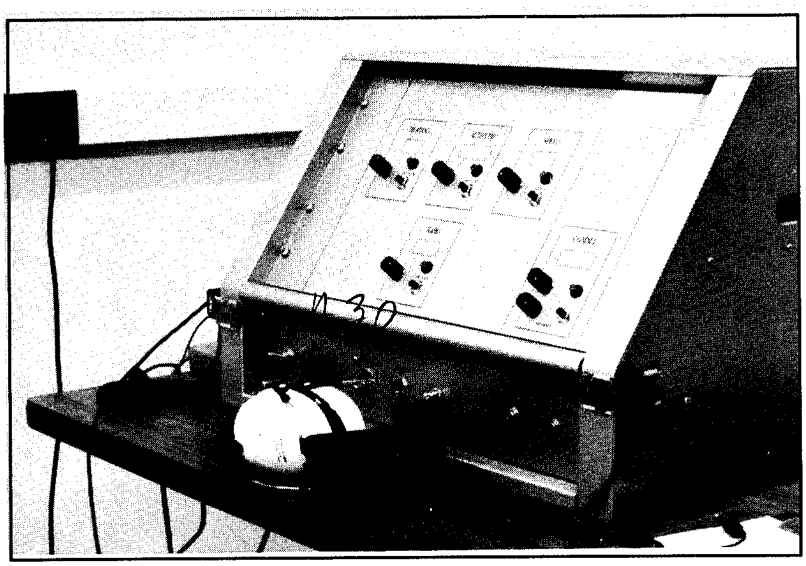


FIGURE 1: Mock-up Mode Control Panel

observed when a new clearance changed only in altitude, but not heading. The order of presentation of the 3 sessions was again counterbalanced across the 24 airline pilots volunteering for the experiment. Each session lasted 45 minutes.

Here is an example of a clearance with 5 pieces of information in the "grouped format:"

> Universal 1642. Reduce speed to two thirty. Fly heading zero four zero. Descend and maintain fifteen thousand. Houston altimeter twenty-nine fifty-two. Contact Houston Center on one thirty-two point twenty-two.

As you can see, heading information is always given sequentially. Here is another clearance at complexity level 5, this time in "restated format:"

> Universal 1642. Climb and maintain one six thousand, that's sixteen thousand. Revised Houston altimeter two niner niner niner. Fly heading two four zero. Increase speed to three one zero. Contact Houston Center on one one niner point seven five, that's one nineteen point seventy-five.

Based on discussions with controllers, we expanded the recent acceptance of restating altitude to include frequency.

RESULTS

Now, let's look at the results. Figure 2 shows percent errors by complexity for each format, summarized over all types of information. Pilots performed remarkably well, especially considering that we were testing unaided recall without a coherent flight scenario. The number of responses collected in each cell ranged from 361 to 2587, and the error rate never exceeded 4.2 percent.

We counted as errors all instances where either readback or setting or both were incorrect, resulting in a total of 224 errors or 2.2 percent. For most errors, both readback and setting were incorrect and identical. In 2 instances, both were incorrect but different. In 13 cases, the readback was wrong, but the setting was correct. There was fortunately only one opposite case, where the readback was correct, but the setting was wrong. Cases where both the readback and setting were omitted were also considered as errors.

Figure 3 shows percent miscommunications summarized over all types of information, by complexity level for each format. Miscommunications include not only errors, but also requests for repeat (regardless of whether the readback or setting were correct), and the 42 instances where pilots set the correct number but omitted the readback that was mandatory in our experiment. In short, miscommunications include anything that taxes air traffic control resources, be it that controllers have to correct readbacks, repeat information, or ask for a readback that they had requested.

As you can see, an increase in complexity does appear to increase the number of miscommunications, especially in the grouped format.

Figures 4 and 5 show percent errors and miscommunications for altitude by complexity for each format. The values ranged from 4,000 to 17,000, in increments of 1,000. The number of responses collected in each cell ranged from 216 to 384.

As you can see, there were very few errors or miscommunications. The possible reasons for this excellent performance with altitude are threefold. First, altitude is arguably the most important piece of information in any clearance. Second, the numbers used in the low-sector en route environment simulated

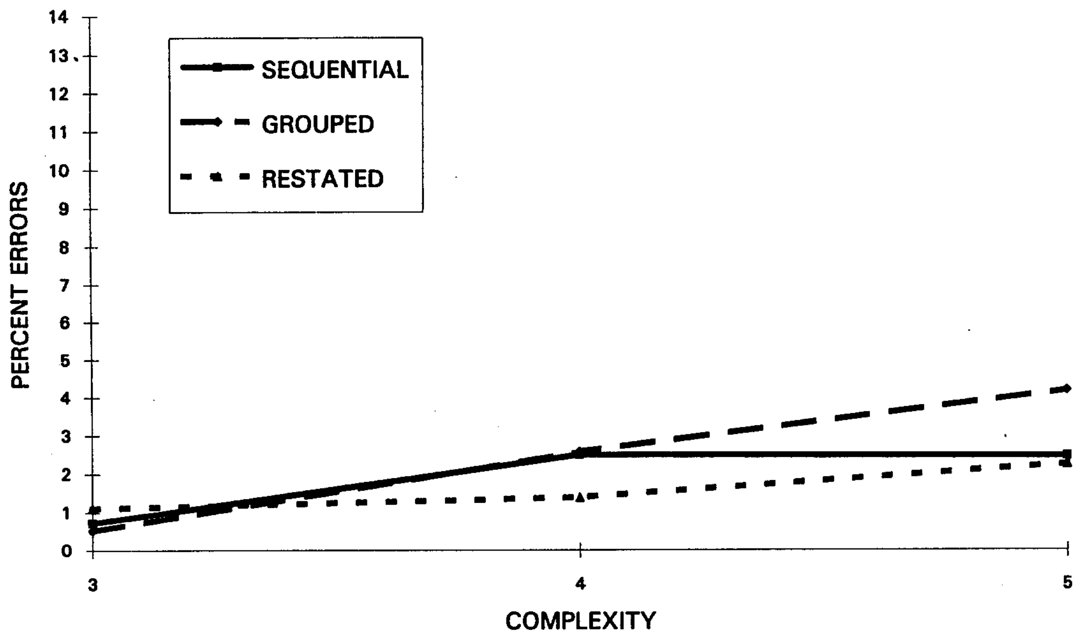


FIGURE 2: Summary of Errors

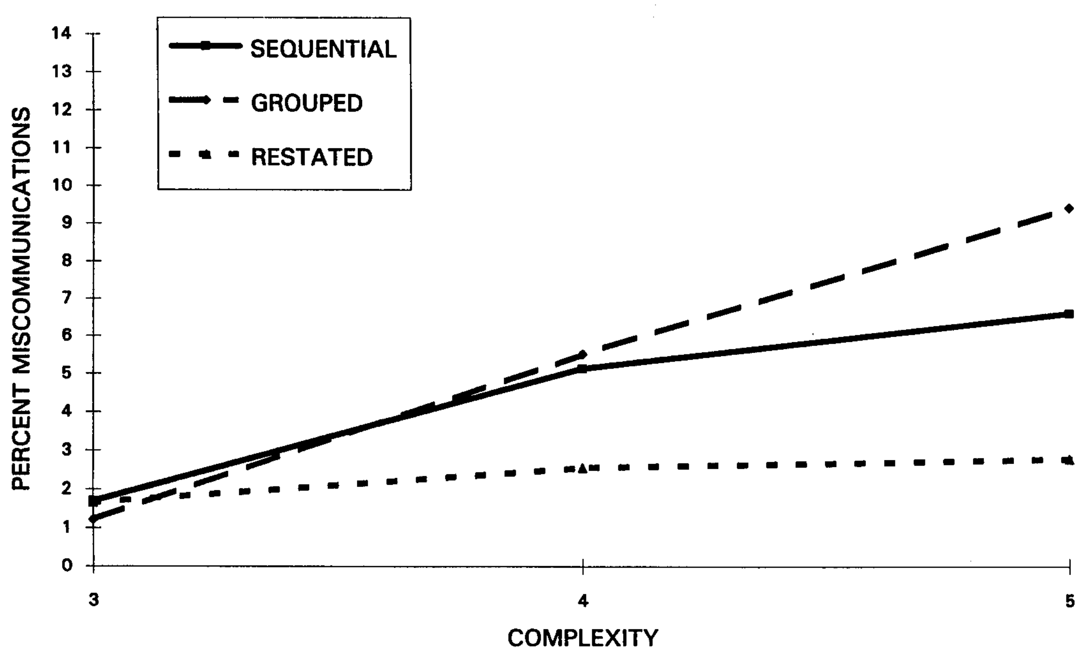


FIGURE 3: Summary of Miscommunications

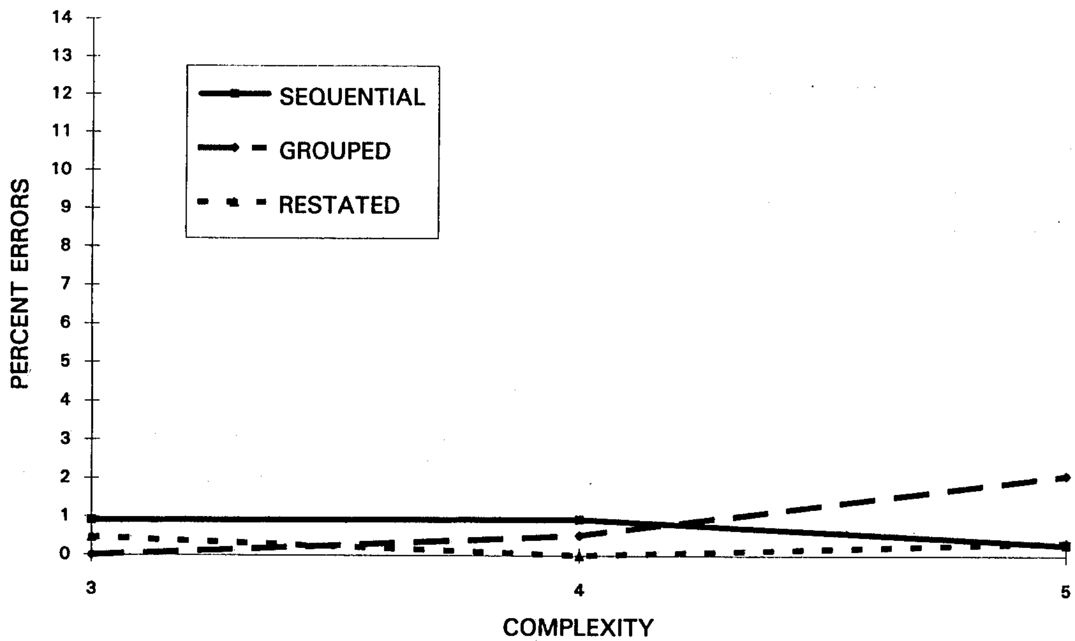


FIGURE 4: Percent Errors for Altitude

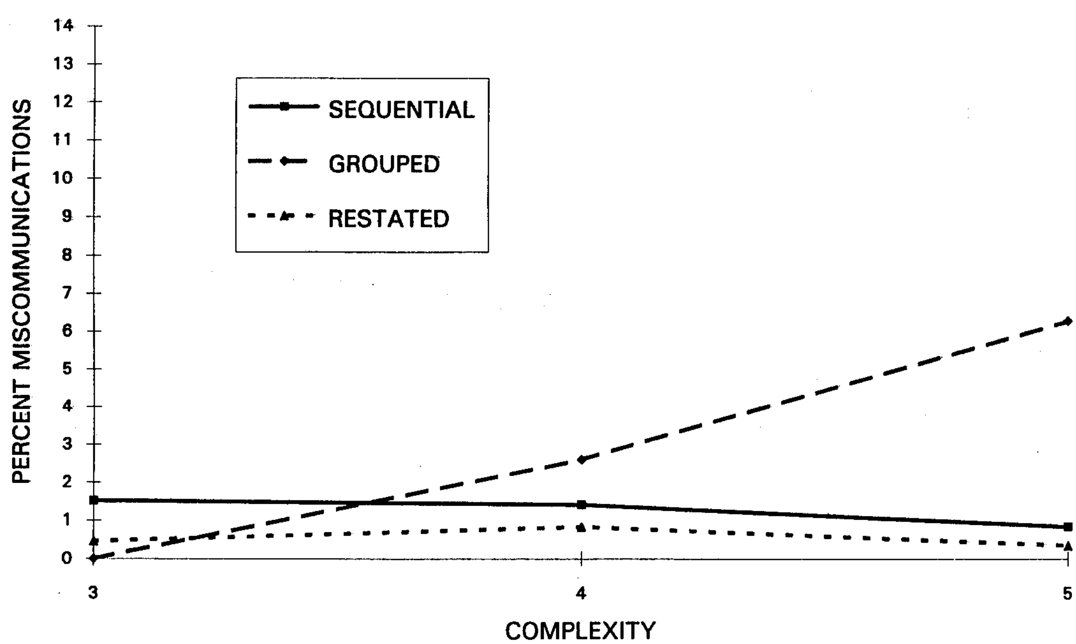


FIGURE 5: Percent Miscommunications for Altitude

in our experiment cover a relatively small range. Moreover, we always gave altitude with a heading and/or a speed, which restricted the set of possible numbers even further (even/odd rule, speed restriction at low altitudes). Third, the number itself, with maximum 2 positions and the "thousand" remaining constant, do not represent a high memory load.

Again, recall of altitude in grouped format appears to deteriorate with increasing complexity. No conclusions regarding the effects of complexity can be drawn for the sequential and restated format due to the small number of errors.

Figures 6 and 7 show percent errors and miscommunications for radio frequencies by complexity for each format. Frequency values ranged from 118.02 to 112.37 and 123.67 to 135.97 in increments of .01. The number of responses collected in each cell ranged from 145 to 287.

A comparison between the 2 figures shows that the apparent stabilization or reduction of errors when going from 4 to 5 pieces of information per clearance is due to an increase in requests for repeat, at least in the grouped format.

Restating frequency does appear to reduce miscommunications. This might however be simply a function of hearing the information twice, regardless of the format.

Figures 8 and 9 show percent errors and miscommunications by complexity for each format for altimeter. As you can see, altimeter was not presented in clearances with only 3 pieces of information. Settings ranged from 29.00 to 31.00 in increments of .01. The number of responses collected in each cell ranged from 192 to 575.

Whereas in grouped format both error and miscommunication rates increase with complexity (the latter to as high as 13.59 percent), in sequential format only miscommunications appear affected, reflecting again the fact that pilots more readily asked for a repeat at the higher complexity level.

Figures 10 and 11 show percent errors and miscommunications for heading, which ranged from 10 to 360 in increments of 10 and was said only in sequential format. The number of responses collected in each cell ranged from 718 to 863. Again, increased complexity appears to result in a modest decrement of recall.

Figures 12 and 13 show percent errors and miscommunications by complexity for each format for speed. The number of responses collected in each cell ranged from 240 to 575.

Ranging from 210 to 310 in increments of 10, speed is the only type of information that appears to show slightly better recall when it is said in grouped format, at least with regard to miscommunications. A possible explanation is that grouping speeds help distinguish them from the always sequential headings, with which they overlap in range. In other words, speed and heading are uniquely encoded in the grouped condition.

An increase in complexity does appear to affect recall, although the percent miscommunications in grouped format level off when going from 4 to 5 pieces of information in a clearance.

CONCLUSION

The following conclusions may be drawn from these data:

1) Restating information appears to improve recall, although the format of the repetition may not matter.
2) A part from the possible coding advantage for speed, the data do not support the widely held opinion that grouping numbers improves recall; indeed, grouping might reduce recall. An objection to this conclusion may be that we were testing the unfamiliar against the familiar in the same test session (mixed design).
3) Presenting more than 3 pieces of information in one clearance may lead to errors or requests for repeat.

These conclusions are supported by pilots' answers to the questionnaire administered both before and after they had experienced the different formats/complexity levels in the experiment. Let's look first at pilots' before and after preferences for format.

For altitude, 13 of the 24 pilots preferred the restated format already before exposure, 16 after. Five pilots preferred grouped before, but only 2 after. Six pilots preferred the sequential format both before and after.

Only 1 pilot wanted frequency restated before exposure, but as many as 11 after. Ten pilots wanted frequency grouped before exposure, but only 6 after. Thirteen pilots preferred frequency sequential before and 8 after exposure.

Most pilots (16) preferred heading in sequential format both before and after the experiment. Three more switched from grouped to sequential after the experiment. Two remained with grouped, 2 with restated, and 1 switched from sequential to restated. Note that in the experiment, heading was consistently said in sequential format.

Despite the possible advantage of grouping speed in our data, only 5 of the original 10 subjects preferring speed grouped remained with it after exposure. One switched to grouped after initially preferring sequential. Restated doubled its adherents from 3 to 6 after the experiment, and sequential added 2 to its original 11. Note that speed was never restated in the experiment.

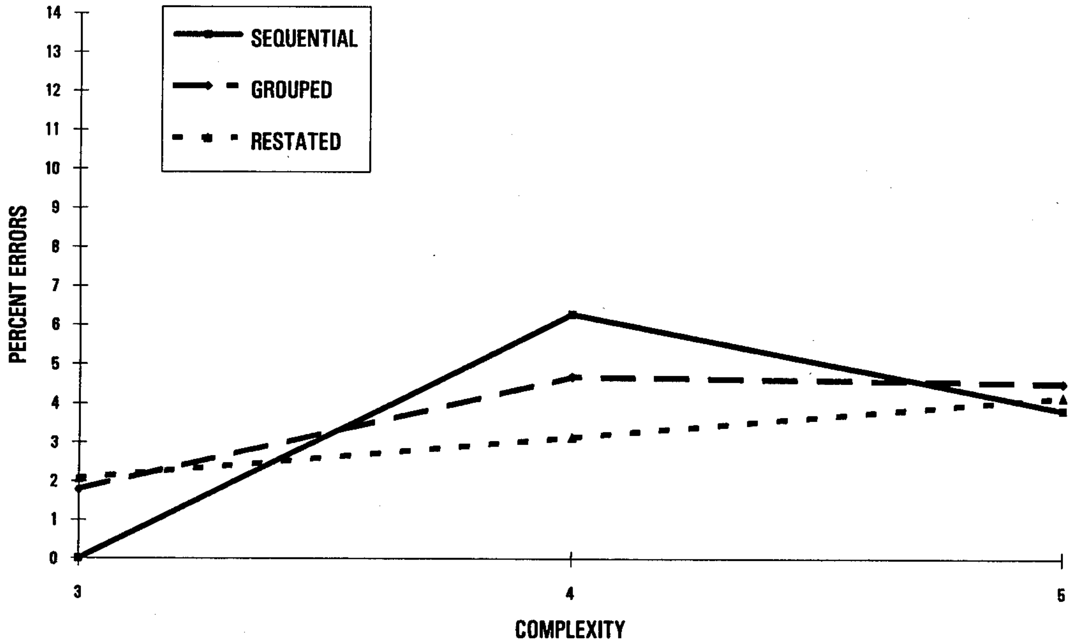


FIGURE 6: Percent Errors for Frequency

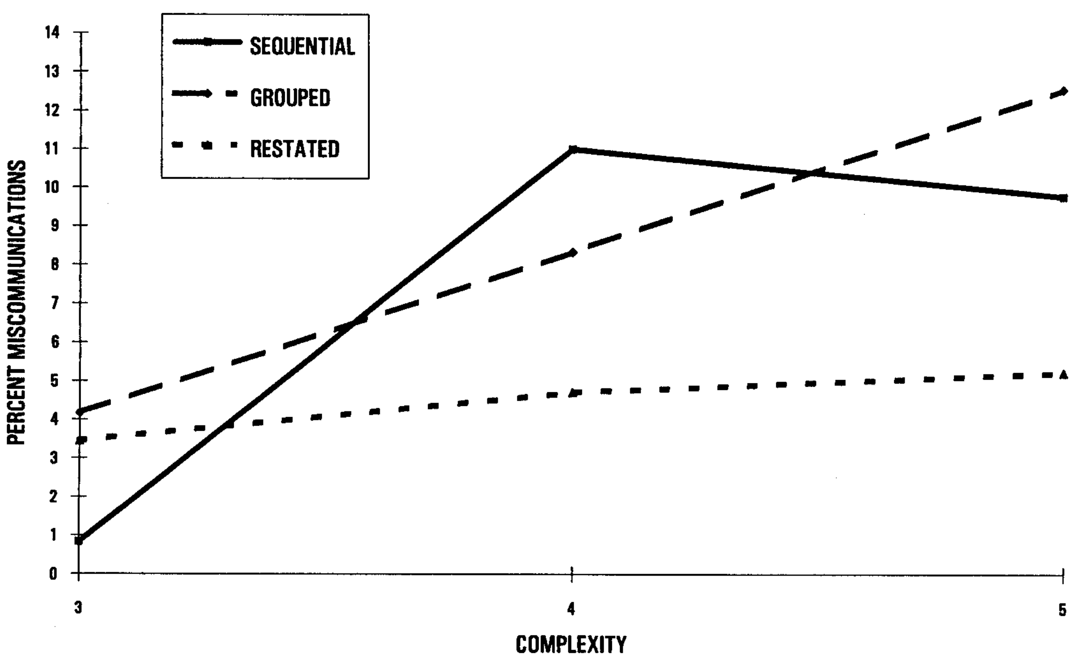


FIGURE 7: Percent Miscommunication Frequency

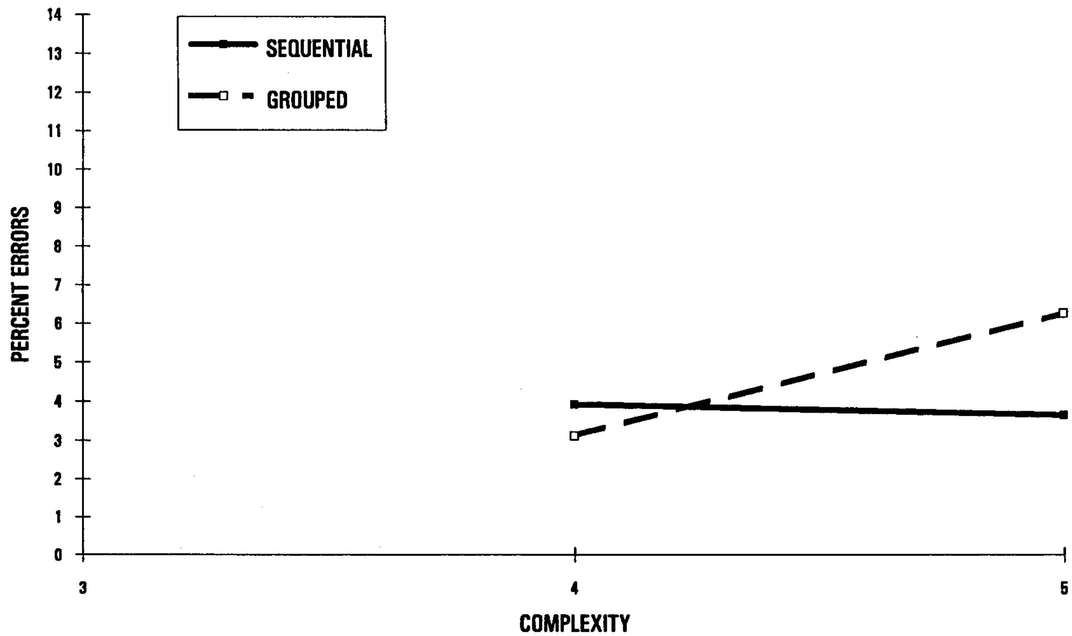


FIGURE 8: Percent Errors for Altimeter

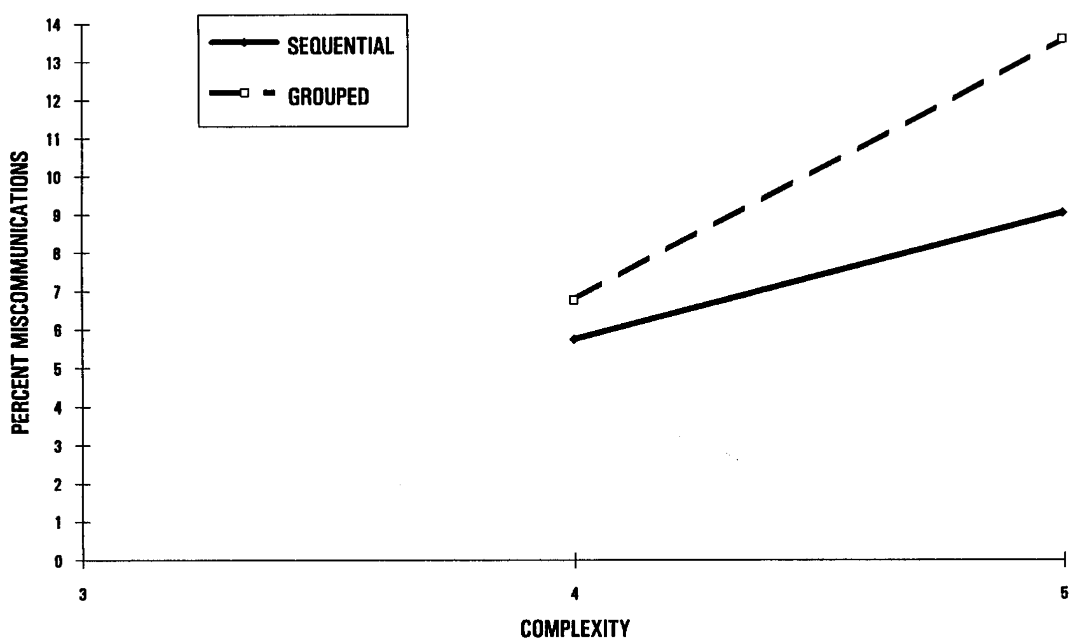


FIGURE 9: Percent Miscommunications for Altimeter

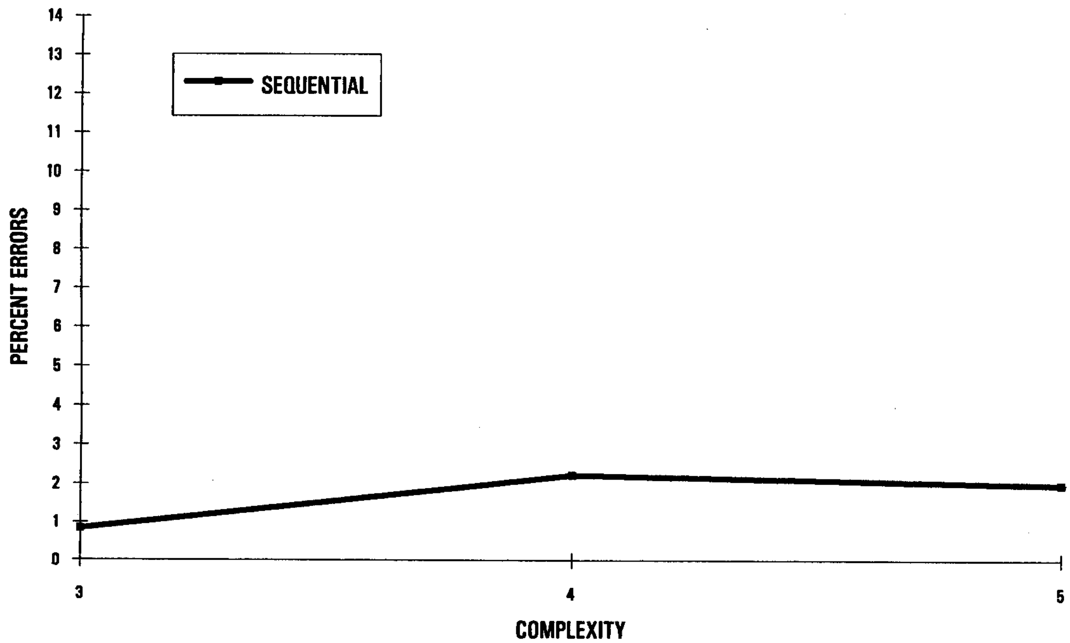


FIGURE 10: Percent Errors for Heading

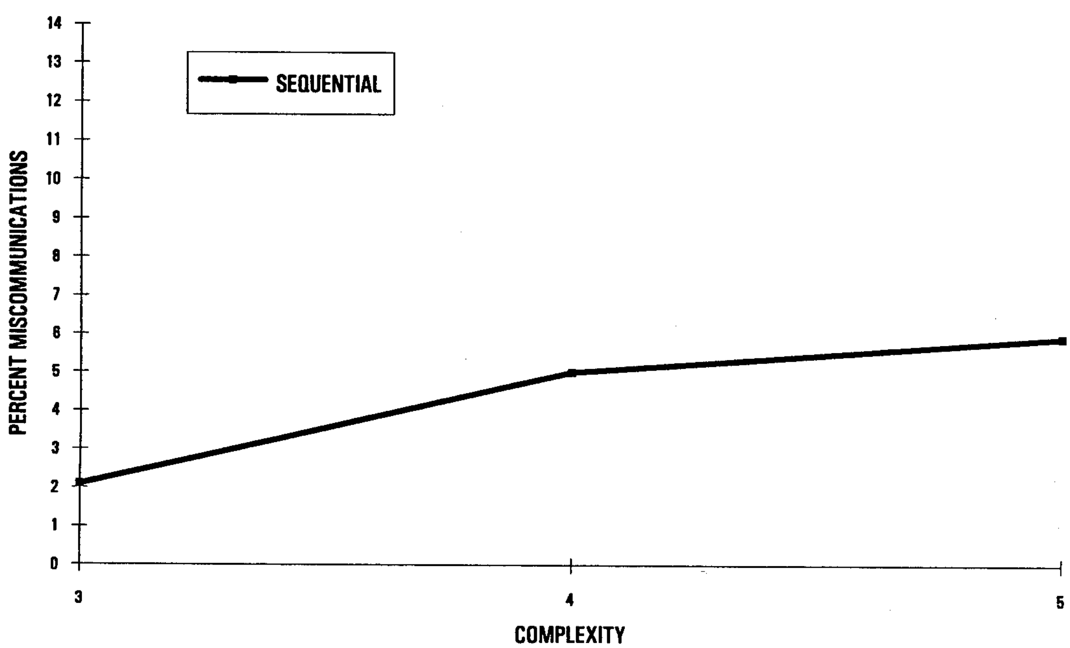


FIGURE 11: Percent Miscommunications for Heading

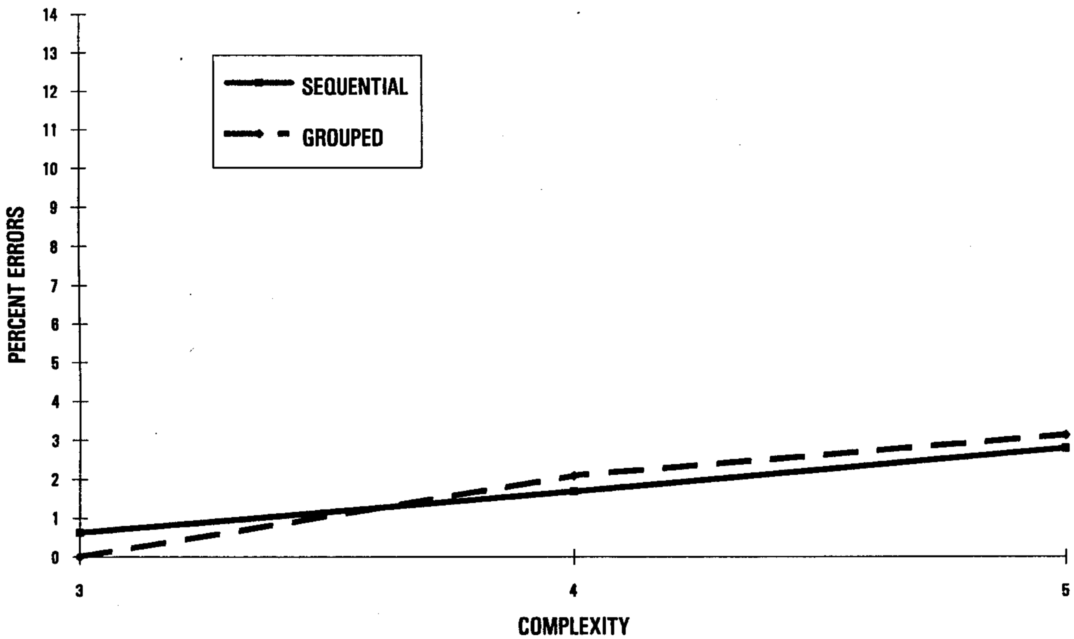


FIGURE 12: Percent Errors for Speed

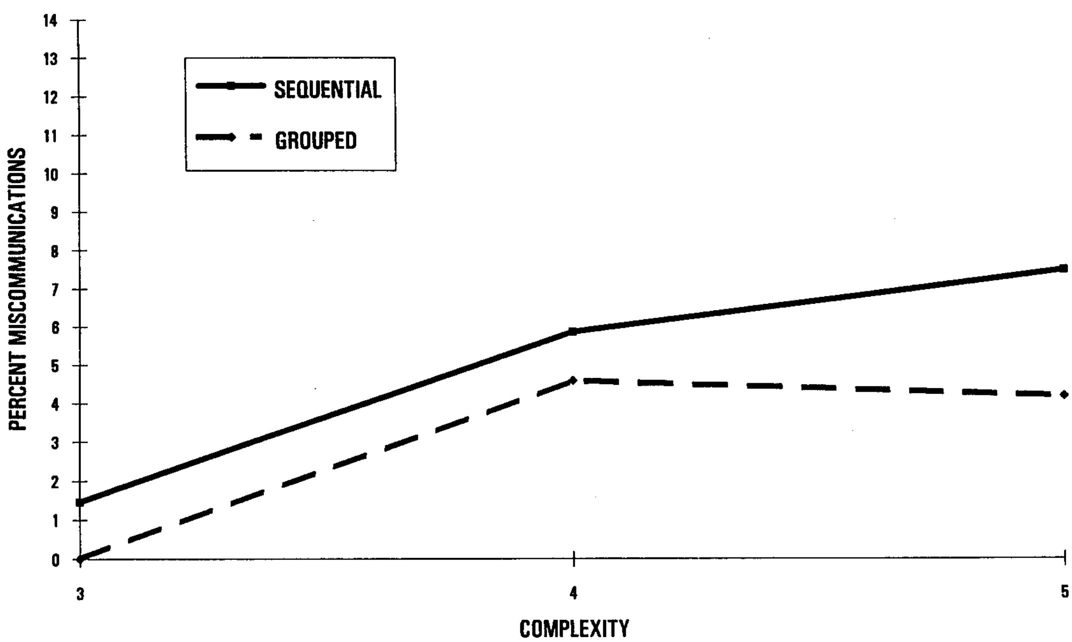


FIGURE 13: Percent Miscommunications for Speed

Altimeter, which was also never restated, received 10 votes for grouped before, and 8 after exposure. One pilot voted for restated before, and 2 after. The number of adherents to the sequential format remained at 13, although there were some losses and gains mainly from and to the grouped format.

As to airline pilots' intuitions about complexity, half recommended a maximum of 3 pieces of information per clearance both before and after the experiment. This number rose to 20 after the experiment (interestingly, 2 of these pilots increased their recommendation from 2 to 3). Three pilots felt that they can handle as many as 4 even after the experiment. One pilot never wants to hear more than 1 piece of information.

In closing, these results do not support the commonly held opinion that presenting numerical ATC information in grouped format helps. They do support, however, the practice of restating information, possibly regardless of format. Moreover, controllers should be advised that presenting more than 3 pieces of numerical information in a single clearance may not save time, but lead to errors or at least requests for repeat. This conclusion is particularly significant considering the fact that our results stem from airline pilots with a minimum of 3,000 hours of experience (although the average was much higher).

We are currently preparing to conduct an experiment that will test whether the effects of format and complexity interact with speech rate. In this experiment, we will test the 3 formats separately (blocked design), which will weaken the argument that testing the old (sequential) against the new (grouped) is an unfair comparison.

REFERENCES

Cardosi, K. M. (1993). *An Analysis of En Route Controller-Pilot Voice Communications.* Springfield, VA: National Technical Information Service.

A TRAINING PERSPECTIVE: ENHANCING TEAM PERFORMANCE THROUGH EFFECTIVE COMMUNICATION

Barbara G. Kanki
NASA Ames Research Center

INTRODUCTION

The presentations given in this workshop are commonly united by their focus on voice communications, and we are here today in order to share our experiences with different methods and metrics of analysis. In addition to studying different aspects and levels of communication, our "subjects of study" are speakers and hearers from different work domains primarily within the aviation system. In addition, our research aims at a variety of research goals. For example, some among us analyze communications because we are trying to understand why an accident occurred; others are concerned with developing equipment and procedures that are better suited to communication requirements. Still others represent a training perspective, and study communications in order to provide guidelines and recommendations for training more effective communicators. This is the perspective I am presenting and much of the work we do at NASA Ames Research Center in crew communication is geared toward customers from airline training departments with the goal of giving trainers useful information for the pilots they train. Through communication analysis we look for patterns and principles that address communication problems and point to effective strategies and skills that can be learned, practiced and assessed.

Conceptual Model for Communication Processes

The conceptual framework in which communication processes play a critical role is depicted below in Figure 1. This is a simple input-output model, in which communication is one type of group process that mediates between a large number of input factors and team performance outcomes. Basically, we are interested in learning about relationships between input factors and process variables that affect performance outcomes such as safety, effectiveness and efficiency of flight operations. We are particularly interested in group processes because these are behavioral patterns that can be identified and described for training purposes; both in terms of patterns that alert us to symptoms of problems, and also in terms of effective patterns that represent successful intervention strategies. In any particular situation, communication may indicate how the team is progressing toward their mission goals. Are there symptoms that a team is having problems? Is there evidence that a team is successfully coping with a situation? Our goal as researchers is to identify patterns that differentiate the 2; lower performing teams from high performing teams.

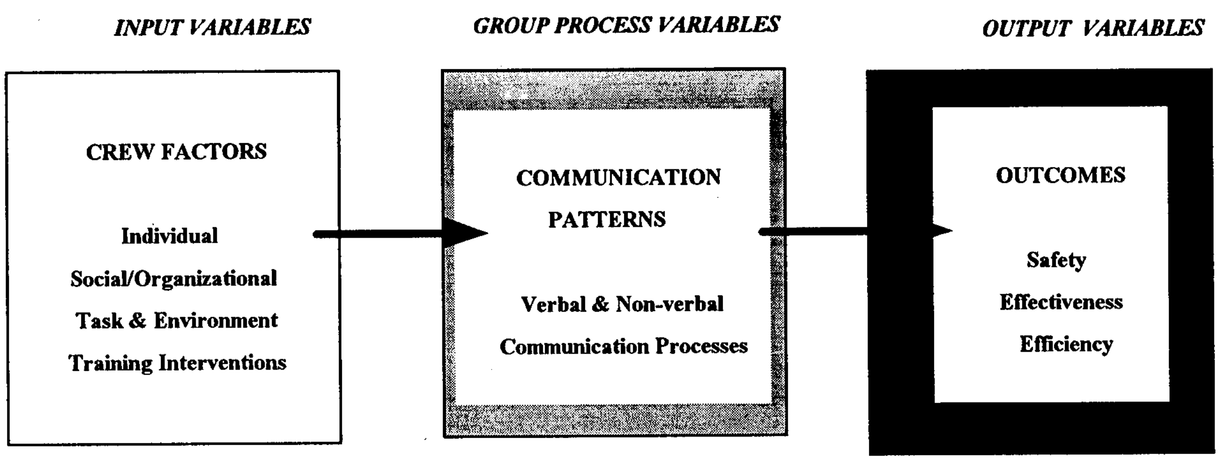


FIGURE 1: Communication patterns may indicate symptoms of problems or strategies for intervention

Assumptions Underlying the Communication Process

Assumption 1: Communication is Interactive.

Our research approach incorporates several key assumptions. First, communication is interactive; that is, communication involves speakers and hearers whose speech acts are both actions and reactions. In short, speech is both active and reactive because communicators are both speakers and hearers (usually in sequence). What one person says may have been generated as a response to the other, and, in turn, it may provoke a response as well. Because of its interactive nature, speech can accomplish far more than statements of reference. For instance, in the aviation context, speech is used to issue commands, acknowledge commands, conduct briefings, and perform standard callouts. In addition, more general functions such as conveying information, asking questions, stating intentions, etc. are other commonly observed performative speech acts. From a team perspective, communication functions such as those stated in the lefthand column in Figure 2, play an important role in the aviation workplace. The column on the right lists the associated consequences when communications fail. Because these types of consequences have been related to aviation accidents and incidents, we are confident that enhanced communication skills can play a role in improving flight safety (Kanki and Palmer, 1993).

COMMUNICATION FUNCTIONS	RELATED CREW COORDINATION PROBLEMS
* **Provide information**	* **Lack or wrong information**
* **Establish interpersonal tone**	* **Interpersonal conflict and tension**
* **Establish predictable behavior**	* **Non-standard, unpredictable behavior**
* **Maintain attention & task monitoring**	* **Loss of vigilance, situation awareness**
* **Management tool**	* **Lack or misdirected leadership**
* **Problem-solving & decision making**	* **Lack of planning, preparedness**

FIGURE 2: Some communication functions and their related problems

Assumption 2: Communication Takes Place in a Context.

Our second assumption is that communication takes place in a context; namely, the physical, social and task environments of the aviation domain. For example, the physical environment includes aspects of the aircraft and flightdeck itself, the communication media used, and features of specific equipment. It also includes physical features of the ambient environment such as noise level, lighting, and competing stimuli. A second context of communication is the social environment. For example, the way we interpret speech is greatly influenced by who says it. Identical words spoken by a captain and a first officer may have very different impacts. Similarly, individuals become known by their own personal styles, and the same words spoken by a "gregarious, impetuous" person may carry a very different meaning when spoken by a "reserved, cautious" person. A third context of communication is the task environment. In the aviation workplace, communications are highly constrained by the structure and standards of the task itself because through it, the standard operating procedures (SOP's), the roles of the operators, as well as the rules of authority carried by each crew position are defined. Embedded within the task environment are the actions of the operators themselves. Although we have been talking primarily about verbal communication so far, communication includes many more aspects of a person's behavior. In the aviation context, there are operator actions, as well as nonverbal signs that carry implicit meaning in the work setting. A gaze toward an instrument, and a move toward a dial may convey more about a person's intentions than words. Much more can be discussed in this domain, but I will leave the topic of communication in the context of nonverbal information to a later talk by Segal (Segal, 1995).

In summary, many aspects of the physical, social, and task environment shape the way communications occur, and we need to understand their effects in order to identify what is constant about communication, and what is free to vary. In analyzing any particular speech stream, our strategy is to systematically trace communication variation back to its source. We need to be able to distinguish whether standard patterns are due to grammar, SOP's, crew position, organizational culture, etc. Similarly, we need to be able to distinguish whether unusual patterns are violations of standards, whether they are acceptable, but ineffective, whether they are exceptionally good examples of crew strategy, or whether they are simply stylistic differences which are unrelated to performance. Once we are able to make these kinds of discriminations, we come closer to identifying specific communication behaviors which can be usefully trained.

From a training perspective, the bottom line is the proven relationship of clearly defined communication behaviors to performance differences. Specifically, we need to know what communication patterns contribute to high levels of crew coordination, and under what conditions they most effectively occur. Are the patterns related to pilot role? Are they related to particular flight phases, or procedures (normal vs. non-normal), routing, weather, emergency conditions, etc. Our primary goal is to develop an information base of effective communication recommendations for training purposes. In addition, we consider related alternatives to enhancing team coordination; for instance, through hardware or procedures design that facilitates the communication process.

METHOD

There are many ways to obtain data, and each method has its own particular strengths and weaknesses. Our program of research attempts to take an integrated approach by recognizing a variety of both field and experimental approaches and using them in concert with each other. For example, although conducting field research is difficult in terms of collecting data in a systematically controlled environment, one gains the natural advantage of high validity. On the other hand, experimental approaches give us better control of variables and often the opportunity to collect large samples efficiently; but these advantages are worthless if results fail to generalize to actual operations.

Field Methods

It is typical for researchers concerned with aviation safety to educate themselves in the problem area by studying accident and incident reports. As providers of field data, both of these are extremely informative resources with high face validity. Accident investigations are particularly valuable because the cockpit voice recorder (CVR) as well as air traffic control tapes provide raw voice data which can be analyzed in a variety of ways. Talks by Veronneau (1995), and Brenner, Mayer & Cash (1995) will address these issues later. (Also see papers by Predmore (1991), and Helmreich (1994) in the Appendix of related publications.) By their very nature, accident analyses are case studies which are limited in number, but allow in-depth treatment. In contrast, incident reports rarely allow deep analysis, but they provide a complementary resource because of the large number and wide variety of cases represented.

The Aviation Safety Reporting System (ASRS), described earlier, is an example of a large-scale incident database. As mentioned, approximately 70% of the first 28,000 reports received (during the first 5 years) were found to be related to communication problems. Because these incident reports represent many locations, many companies, a great variety of equipment and work conditions, the database provides a system-wide perspective on communication issues. On the other hand, the data is limited because it is difficult to ensure that the data has been collected in a standard way. Because these are voluntary self-reports, the data is already "interpreted" to some extent, and may represent biased perspectives. Nevertheless, as described earlier in Connell's presentation (Connell, 1995), the ASRS database has been very useful in identifying pilot communication errors, air traffic controller (ATC) communication errors, as well as insight into the pilot-ATC communication process (e.g., hearback-readback problems).

Finally, I would like to mention the observational approach to field data collection. In this method, actual operations are directly watched by observers (e.g., jumpseating, online behavior coding), or indirectly recorded via audiovisual means. Obviously, the use of actual operations preserves face validity, but the sampling of activities is a critical choice the researcher must make. If the best choices are made, the critical behaviors will be highlighted in an unbiased sample.

Online observations are constrained by the work environment (physical space, safety, rules, etc.). Many operations are conducted in spaces where observer access is limited or where an observer's presence would be highly intrusive; and some high-risk operations are conducted in spaces where observers are not permitted for safety reasons. Recording observations may solve some of these problems but the use of tapes has its own difficulties. Audio and video recording requires legal, logistical, and physical access to the behaviors of interest, and such access is not always easy to obtain. Finally there are some operations which are so complex or which are so remote, that the means for obtaining all aspects of the operations is very difficult. Here, we must rely on obtaining access to an existing communication system which links all participants should one be available.

Experimental Methods

The obvious payoff of using experimental methods is that complex operations can be controlled to some extent. Unfortunately, what is gained in experimental control often results in a loss of operational realism. Therefore the researcher must carefully find the point on the continuum of realism (from laboratory to full mission simulation) which provides enough control to conduct a meaningful experiment. For example, if the research question focuses on an individual operator's response, data collection on an individual level in the laboratory or part-task simulator may be a suitable choice. However, if the research focuses on system responses, and involves the way in which multiple crew members or multiple teams work together, a more realistic full mission simulation may be needed. This decision is also affected by whether the research questions are highly focused or whether it is a more exploratory investigation. If the research process has just begun, it may be too early to restrict the research design to a simple hypothesis testing paradigm.

Even within the full mission simulation paradigm, there are many choices to be made by the researcher. There is a certain degree of control imposed by the simulator itself; that is, pilot actions are constrained by the actual machinery and research environment. Still, an appropriate scenario design is needed in order to (1) elicit the behaviors you want to study, and (2) enable the unconfounded comparison of critical behaviors. Specific design decisions must be based on the particular research question. Are you interested in an individual or system response? Are you interested in very specific actions or are you exploring many alternative actions? In many of our simulation studies, we included pilot-ATC communications in the scenario, since its omission would be highly unrealistic. However, we simply scripted the controllers' communications in order to maintain greater experimental control. Since we were focusing on pilot actions only, this decision made sense. But if we were to design a simulation focusing on both controllers and pilots, we would have to trade-off experimental control in order to elicit the cross-team communications we want to study. In general, the scenario design is the key to answering your research questions. It must be controlled enough that a systematic analysis is possible, but it cannot forsake operational realism and the opportunity for behavior variations to emerge. It is a fine line which requires a careful decision.

The above represents only a few issues in designing a research method for studying communication processes. There are 2 main lessons to learn; (1) there are tradeoffs for every method, and (2) the most appropriate choice is dependent on your particular research question. What I would like to do is leave these discussions as open topics, especially since our next talk by Veinott and Irwin will explore many of these issues in greater depth (Veinott & Irwin, 1995). I will also leave it to them to present actual examples of some of the choices we have made in addressing particular research questions.

CONCLUSION

In conclusion, I simply want to reiterate the overall objective of this workshop; namely that we have many experiences represented here, many research methods and objectives. While I'm sure you have all enjoyed a measure of success from your current practices and the trial-and-error learning that you've been experiencing over the years, it would be nice to circumvent some of this reinvention of methods by assessing our methods and objectives, and matching them up in innovative and perhaps more effective ways. If we can share our experiences, we can learn about which matches work well and which don't. Finally, with the understanding that different approaches yield different kinds of research answers, we can expand our research by integrating multiple methods. Your research methods integrated with mine may give us both better, more comprehensive research answers.

REFERENCES

Brenner, M. Mayer, D., & Cash, J. (1995). *Speech analysis in Russia.* In B. G. Kanki & O. V. Prinzo (eds), Proceedings of the Methods and Metrics of Voice Communications Workshop.

Connell, L. (1995). *Pilot and controller communications issues.* In B. G. Kanki & O. V. Prinzo (eds), Proceedings of the Methods and Metrics of Voice Communications Workshop.

Helmreich, R. L. (1994). Anatomy of a system accident: The crash of Avianca Flight 052. *Intl. J. Aviation Psych,* 4 (3), 285-291.

Kanki, B. G. & Palmer, M. T. (1993). Communication and crew resource management. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich, (eds.) *Cockpit Resource Management,* San Diego: Academic Press, pp. 99-136.

Predmore, S. (1991). *Microcoding of cockpit communications in accident analyses: Crew coordination in the United Airlines Flight 232 accident.* In R. S. Jensen (Ed.), Proceedings of the Sixth International Symposium on Aviation Psychology (pp. 350-355). Columbus: Ohio State University.

Segal, L. (1995). *Speech in the context of nonverbal activity.* In B. G. Kanki & O. V. Prinzo (eds), Proceedings of the Methods and Metrics of Voice Communications Workshop.

Veinott, E. S. & Irwin, C. M. (1995). *Communication methodology issues in aviation simulation research.* In B. G. Kanki & O. V. Prinzo (eds), Proceedings of the Methods and Metrics of Voice Communications Workshop.

COMMUNICATION METHODOLOGY ISSUES IN AVIATION SIMULATION RESEARCH

Elizabeth S. Veinott and Cheryl M. Irwin
San Jose State University Foundation

INTRODUCTION

This paper will address several communication methodology issues encountered during the analysis of data from an aviation simulation study. We will review the data collection, transcription and coding of communications and briefly describe the data analysis and results of a study comparing the communications of pilots in the automated MD-88 with those of pilots in the traditional DC-9 aircraft. By highlighting aspects of the methodology involved, we can address specific tradeoffs that have been encountered involving the simulation scenario, transcription of the videotaped data, communication coding, selection of dependent variables, and level of analysis. We present these issues not as a means of solving methodological dilemmas, but more as a means of generating discussion about communication research in general and the special considerations for communication research in the aviation simulation environment. A complete discussion of the original simulation study and analyses can be found in Weiner, Chidester, Kanki, Palmer, Curry & Gregorich (1991), and fuller descriptions of the associated communication analyses are in Kanki, Veinott, Irwin, Jobe & Wiener (in prep).

METHOD

Twelve DC-9 and 10 MD-88 two-person crews from a major U.S. air carrier participated in a full-mission simulation. The scenario was designed to include both low- and high-workload periods. During the flight from Atlanta, GA to Columbia, SC, the crew had to execute a missed approach at Columbia due to bad weather. The flight was diverted and eventually landed in Charlotte, NC. Following missed approach, which began the high workload portion of the flight, the crew had to select an alternate and perform system malfunction checklists due to the failure of a constant speed drive generator. Total flight time was approximately 80 minutes (Figure 1).

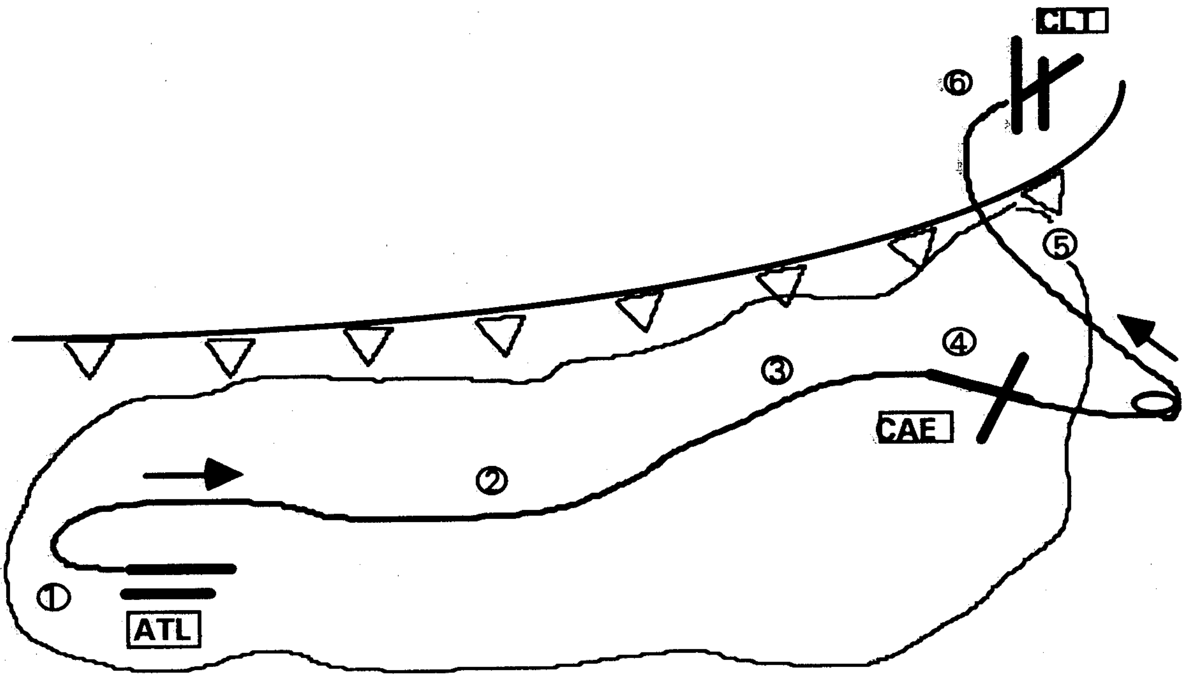
Prior to the flight, each pilot completed a demographics questionnaire. The simulation was videotaped from 2 camera views, giving an over-the-pilots' shoulders view of the flightdeck, as well as a separate view of the controls. Each pilot's voice was recorded on a separate audio channel. An onboard observer rated each crew's performance with respect to coordination, task management, and aircraft handling. At the end of the experiment, pilots rated their own workload and performance. Two NASA expert observers reviewed the videotapes and rated pilot errors. Overall, there were no differences in performance between the 2 aircraft. There was a significant difference in total flight time, with MD-88 crews averaging a longer flight time in the Normal phase. Also, MD-88 pilots' self-reported workload ratings were higher than those of DC-9 pilots.

Videotapes of the flights were transcribed from push-back in Atlanta to touchdown in Charlotte. All checklists and air traffic control (ATC) communications were transcribed. Speaker, and start and end times of the speech were recorded. The verbatim transcripts reflected the actual occurrence of speaker turns, and were subsequently unitized for coding purposes. Four independent coders trained to reliability (Cohen's kappa > .75) on the use of 14 speech codes (Bales, 1950; Foushee and Manos, 1981) and coded the communications for speech category and initiation-response information using the transcripts and video tapes (Figure 2).

For the purposes of analyzing the initiation-response patterns, the above codes were collapsed into 4 initiation and 3 response categories. Initiations consisted of commands, observations, questions, and dysfluencies; and responses consisted of replies, acknowledgments and no responses. Initiations and responses were mapped into a 4 by 3 matrix for each of 2 directions of speech: captain as initiator, first officer as responder; and first officer as initiator, captain as responder. The following example is a matrix of captains' initiations and first officers' responses (Figure 3).

Dependent Measures

Because of the difference in flight time mentioned earlier, 2 dependent measures that provide qualitatively different information were used: time-adjusted frequency and proportion of total communication. Time-adjusted frequency (TAF) adjusts the total number of speech acts in each category by flight time or opportunity to speak. TAF's control for flight time differences and give a density of communication to enable communications per minute comparisons. A second dependent variable, proportion of total communication (PTC), controls for differences in total communication and measures the distribution of



1. Departure
2. Cruise to CAE
3. Initial approach to CAE
4. Approach/Missed approach
5. Cruise to Alternate
6. Landing

ATL	Atlanta, GA
CAE	Columbia, SC
CLT	Charlotte, NC

FIGURE 1: Automation Simulation Scenario

---

14 Speech Categories

- Command
- Suggestion
- Observation
- Question
- Statement of Intent
- Agreement
- Disagreement
- Acknowledgment
- Answer
- Response Uncertainty
- Tension Release
- Repeat
- Checklist
- Dysfluency

FIGURE 2: Speech Categories

CA -> FO	**Reply**	**Ack.**	**No Resp.**	**Row%**
Command	571	621	461	24%
Observation	1573	963	1816	63%
Question	488	22	80	8.5%
Dysfluency	115	21	191	4.5%
Column %	40%	23%	37%	6922

FIGURE 3: Example Initiation-Response Matrix

Flight Time: Crew A: 60 minutes Crew B: 60 minutes

	Commands	Observations	Questions	Total
Crew A:	$\boxed{20}$	50	30	100
Crew B:	$\boxed{20}$	130	50	200

Commands: Calculating time-adjusted frequency and proportions.

	Time-adjusted frequency	Proportion of total communication
Crew A:	.33	.20
Crew B:	.33	.10

FIGURE 4: Time-adjusted Frequency and Proportion of Total Communication

communication relative to one's own speech. TAF's answer questions regarding frequency, while PTC's address questions about patterns and distribution. The example in Figure 4 clarifies this distinction.

In this example each crew made 20 commands; consequently, a comparison of raw frequencies would show no difference across crews. Now consider the 2 dependent measures described above. Since there is no difference in flight time, the time-adjusted frequencies produce essentially the same comparison as the raw frequencies. Crew A has the same time-adjusted frequency as Crew B (.33 vs. .33) even though overall Crew B has twice as many communications as Crew A. However, for the proportion of total communication, Crew A has a higher proportion of commands than Crew B (.20 vs. .10). Therefore, although the 2 crews have the same frequency of commands, the 2 crews differ in their distribution of commands relative to their total communications.

Analyses

Two methods of data analysis were used to explore these data. The first was log-linear analysis which is a method for comparing 2-way matrices across crews and has been used successfully in past aviation research (Kanki, Lozito, and Foushee, 1989). However, due to the large variability within each aircraft type, relatively few patterns emerged that distinguished between the 2 aircraft. A second approach analyzed group differences using a mixed-factorial ANOVA with Speaker (CA, FO), by Aircraft (DC-9, MD-88),and by Phase (normal, abnormal) for each communication category. Speaker was nested within aircraft so that speaker comparisons were only conducted within each aircraft. Phase was the repeated variable. Since we were mainly interested in differences between MD-88 and DC-9 aircraft, the results focus on aircraft differences and how they were affected by phase and speaker.

Overall Analyses. In general, captains talked more than first officers and there was more communication during the abnormal phase than the normal phase. The distribution of communications in the 2 aircraft differed in the following categories: commands and replies. Overall, DC-9 crews had a higher proportion of commands than MD-88 crews. The aircraft by phase interaction for commands was significant, with DC-9 crews having a higher proportion of commands than MD-88 crews only during the abnormal phase. The aircraft by phase interaction was significant for replies with MD-88 crews having a higher proportion of replies than DC-9 crews only in the normal phase. The time-adjusted frequency data showed that MD-88 crews communicated more than DC-9 crews. Main effects for aircraft show that MD-88 crews asked more questions, made more observations, and gave more replies.

Question study. The differences between MD-88 and DC-9 crews in time-adjusted frequencies were found in observations, questions and replies. These 3 categories seem to be information-providing categories as opposed to action categories, which suggest that information-transfer processes are different in the 2 aircraft. In an effort to better understand these differences, questions were selected for further analysis. Questions are an ideal approach to the study of information-transfer because of their direct nature. A question is an intentional disclosure of an information deficit, while an answer is a public means of providing the information requested. One thousand one hundred and seventy 3 questions from the original data set were coded at 3 levels: function (information-seeking, verification); topic (navigation, book, system, other); and answer (yes/no, information, no answer). The design was again a 2-speaker (nested within aircraft) by 2-aircraft by 2-phase (repeated measure) mixed-factorial ANOVA.

The main findings for this study were higher time-adjusted frequencies for MD-88 crews in the information-seeking, navigation, system, information answer, and no answer categories. A significant aircraft by phase interaction showed that MD-88 crews had a higher time-adjusted frequency than DC-9 crews only during the abnormal phase for navigation and no answer. Information flow was not more interrupted in one aircraft type as compared to the other. In general, questions were answered in the next speaker turn more than 96% of the time for crews in both aircraft.

Integrating Nonverbal Communication

Jobe (1994) analyzed verbal communication and discrete system and navigation control actions during a 10-minute high-workload flight segment to investigate how workload management and pilot roles vary as a function of automation. Integrating verbal communication and actions provides 2 different perspectives of what is occurring on the flight deck. The frequency of verbal communication for each aircraft did not significantly differ during this flight segment. For control actions, the aircraft by speaker by behavior ANOVA was significant. Simple effects analyses revealed that DC-9 crews exhibited more system control actions than MD-88 crews and MD-88 crews exhibited more navigation control actions than DC-9 crews. Pilot roles also differed across aircraft. DC-9 captains exhibited more system control actions than first officers, while MD-88 captains and first officers had approximately the same amount of system actions. However, for navigation DC-9 first officers had more navigation control actions than DC-9 captains and again MD-88 captains and first officers had a relatively equal distribution of navigation control actions. Jobe concluded that DC-9 crews' roles seem to be consistent with traditional flight deck structure with the captain flying and the first officer navigating. MD-88 crews seem to redistribute these behaviors or compensate especially during high workload.

Tradeoffs

We would like to discuss some methodological issues we came across through the course of data collection and analysis. The 5 areas we would like to focus on are: simulation scenario, transcription, coding, selection of dependent variables, and analyses.

Simulation Scenario: Experimental Control vs. Generalizability. In an experimental setting, there is always the tradeoff between experimental control and the generalizability of findings. With this simulation, we were able to gain experimental control, yet maintain a realistic environment. The high fidelity, full-motion simulator allowed us to offer a very realistic flight scenario to the pilots and effectively reproduce the scenario for all 22 crews. The scenario and air traffic control (ATC) transmissions were also scripted in order to maintain more experimental control.

The very nature of simulation, however, limits generalizability. We are restricted to 1 scenario, a small number of subjects (22 crews), a single airline and, in this case, 2 aircraft types. It may be difficult to generalize findings beyond the scope of these constraints and will require replication in another simulation or field study before recommendations can be made.

Transcription: Time Spent vs. Level of Detail. One of the first decisions that needs to be made involves whether to transcribe the videos, and if so, at what level of detail. Variables that contribute to the net cost of the transcription process included amount of tape to be transcribed, number of speakers and their rate of speaking, transcriber expertise, level of detail of the transcript, and the researcher's goals.

The average duration of a flight from push-back to touchdown was roughly 80 minutes: 50 minutes in the normal phase and 30 minutes in the abnormal phase. The length of time it took to transcribe 1 minute of tape during the normal phase was about 15 minutes. This time increased to 30 minutes for every minute of tape during the abnormal phase when the pilots were speaking more rapidly and the ATC transmissions were more frequent.

We have considered having people with flight experience transcribe the tapes. Of course, the accuracy of the transcripts would have been increased somewhat, and might have taken less time to complete, but the cost would have been prohibitive. The "transcriptionisms" below are examples of errors we might have avoided by having "experts" do the transcribing. These errors were infrequent, and easily repaired. Since the task of transcribing is quite tedious, and it may not have kept the attention of an expert through the duration of the process, we chose to have the tapes transcribed by graduate research assistants.

Increasing the level of detail in the transcripts requires an additional commitment of time and money. In past simulations, the level of transcription detail has varied. Some transcriptions omitted start and end times of the speakers' speech, ATC transmissions or normal checklist procedures. This level of detail was not deemed necessary for the analyses at the time, but later, researchers with different questions had to go back through the tapes to transcribe the missing portions. Decisions such as these play a large part in the

AS TRANSCRIBED	AS SPOKEN
No bumpers here	New numbers here
9-16 on the barrel altimeter	9-16 on the baro altimeter
Oil pressure low light eliminated	Oil pressure low light illuminated
Balance cross beams closed	Balance cross-feeds closed
Crossed-eye locket	Cross-tie lock out
Were gonna set our brakes	We're going to center tank
You want to depress to 160	You want to request 160
Airspeed bumps	Airspeed bugs
Aphids	TIS

FIGURE 5: Transcriptionisms

tradeoffs between immediate research goals and future research possibilities. Often times, researchers make choices with only immediate research goals in mind rather than working to establish a thorough archival database suitable for a variety of future investigations.

Coding: Reliability vs. Interpretability. There are tradeoffs for achieving acceptable levels of reliability depending on the number of codes and the coding method employed (i.e., individual coders, pairs of coders or team coding). In this study, we used 4 independent coders trained to reliability (Cohen's kappa > .75) on the 14 codes. Though some coding categories were easier to code reliably in a short period of time, training for all 14 categories took about 9 months. Though the training could have been completed more quickly by using consensus coding or fewer individual coders, the training of the 4 individuals helped to establish some validity of the coding scheme. After 9 months of training, the result is a set of coding rules that facilitate the interpretation of the data.

It is also more difficult to achieve acceptable levels of reliability for higher-level conceptual codes that require more judgment on the part of the coder, such as management or problem-solving codes. In these cases, coders and researchers often invent quick-fix rules that bolster reliability at a possible cost to interpretability. Another means of achieving reliabilities more quickly is to reduce the number of categories in the coding scheme. This can be done by collapsing similar categories into broader categories. In most cases this approach saves time, but interpreting group differences in 'catch-all' categories can be quite difficult.

Dependent Variables: Sampled Data vs. Total Data. Very large communication data sets can be rather unwieldy, so researchers may sample the data. Two approaches to data-sampling in simulation research are time-bound and event-bound. The time-bound approach uses set time segments for coding communications (e.g. the 10 minutes after takeoff). This method saves coding and transcription time, but may limit the number of independent variables that can be investigated. However, this approach does not guarantee that the same events will occur for each crew during that time. If certain events in the simulation are important to answer the research questions, then the event-bound approach is preferable.

Each of these methods impacts one's choice of dependent variables. A researcher may choose to adjust the raw frequencies for time, total number of speech acts, or analyze the data sequentially. Time-bound data control for differences in total time, so raw frequency of speech may be used as a dependent variable. Event-bound sampling often produces differences in total time, so the data must be adjusted for time. An alternative to time-adjustment is proportion of total speech. This dependent variable reflects the relative distribution of speech in each category and controls for differences in total speech. A final approach to the analysis of sampled data is to use sequential information. These sequences can include first order: speech category; second order: initiation-response sequence, or higher order.

Analyses: Meaningful Data vs. Statistical Power. The final tradeoff is between meaningful data and statistical power. This tradeoff affects what type of data a researcher collects and the analyses that can be conducted.

A researcher's investigative approach determines what is meaningful data. Using our data set as an example, a communication researcher who adopts a theoretical approach may use all 10,000 individual speech acts as the data set in order to investigate general issues in communication. Conversely, the human factors researcher who adopts an operational approach may use the sample of 22 crews to investigate issues at the group level, such as pilot role, aircraft type or workload.

Once a determination is made as to the appropriate level of data analysis, problems may arise due to a small number of cases or data points. Some data sets, such as reports from the National Transportation Safety Board and Aviation Safety Reporting System, provide highly relevant information regarding aspects of crew coordination and communication, but these data are limited to a handful of instances and yield data for which only descriptive statistics are appropriate. With simulation data the number of communications for each subject is much larger, thus allowing for a variety of statistical analyses.

Even with large data sets, small representation in categories of interest can prove to be challenging. Two issue arise here: first, what is the minimum frequency required for a category to be statistically analyzable? Second, how does one compensate for the fact that a small frequency category may be very important, but hard to analyze using traditional statistical methods? One solution is to collapse small categories with other similar categories for the purpose of analysis.

CONCLUSION

This paper has presented a few of the tradeoffs we have encountered in our research process. Some tradeoffs address general methodological issues that are applicable in many research settings, while others uniquely impact the aviation simulation environment. Tradeoffs occur at all levels of the research process from data collection and coding to statistical analysis. Ultimately, the researcher's goals determine most of what can be done with a data set and every decision along the way can impact future usability of the data set. By discussing an approach, the decisions that led to that approach, and the lessons learned, we have attempted to provide a vantage point for methodology comparison and discussion.

REFERENCES

Bales, R.F. (1950). *Interaction process analysis: Theory, research, and application.* Reading, MA: Addison-Wesley.

Foushee, H.C. & Manos, K.L. (1981). *Information transfer within the cockpit: Problems in intracockpit communication.* In C.E. Billings and E.S. Cheaney (Eds.), Information transfer problems in aviation systems (TP 1875). Moffett Field, CA: NASA-Ames Research Center.

Jobe, K. (1994). *The effect of automation on pilot role and workload distribution.* Unpublished master's thesis. San Jose, CA: San Jose State University.

Kanki, B. G., Lozito, S., & Foushee, H.C. (1989). Communication indices of crew coordination. *Aviation, Space and Environmental Medicine.* 60, 56-60.

Kanki, B. G., Veinott, E. S., Irwin, C. M., Jobe, K., & Wiener, E.L. (In prep). *The impact of cockpit automation on crew coordination and communication: II. Analyses of speech and nonverbal activity across pilot roles and phases of flight.* (NASA Technical Report). Moffett Field, CA: NASA-Ames Research Center.

Wiener, E.L., Chidester, T.R., Kanki, B.G., Palmer, E.A., Curry, R.E., & Gregorich, S.E. (1991). *The impact of cockpit automation on crew coordination and communication: I. Overview, LOFT evaluations, error severity, and questionnaire data.* (NASA Contractor Report 177587). Moffett Field, CA: NASA-Ames Research Center.

SPEECH IN THE CONTEXT OF NONVERBAL ACTIVITY

Leon Segal
Western Aerospace Laboratories, Inc.

INTRODUCTION

And now for something completely different.... Here I am in a speech communication conference talking about nonverbal communication; so where's the connection? What I attempt to do in my work is take a more "holistic" view of the process of communication, look at *all* the elements that contribute to it, and try to arrive at a greater understanding of crew communication and coordination from a somewhat different perspective.

Anybody who has tried to build a natural language parser—or some kind of "artificial intelligence" program that can generate and comprehend speech—well, you know how incredibly complicated this task can be. And the question is, rather than look at, say, where speech fails, let us approach this issue from the other side: why does communication work? How do we manage to actually make this extremely complex exchange of symbolic sounds called "speech communication" work?

I'd like to start out by *telling* you a joke. In fact, I will tell you the speech elements which are integral—and essential—for the joke. The context is an airline cockpit. Captain says: "Sorry to interrupt you, folks, but we've just had a report of some turbulence ahead, so please stay in your seats a little while. Ready?... One, two, three! Well folks, guess we're through the worst of it and... Oh! Wait.. Looks like we're coming in to some more turbulence!.."

So, did you "get" the joke? Well, the point is, this joke relies on the *interaction* between words and activity to be funny. In fact, most cartoon humor indeed relies on the visual—the non-verbal—to create a humorous situation. If you were to see this cartoon, you'd get a different perspective on what is taking place. Essentially, after the captain warns the passengers of the turbulence, he looks over at the copilot, says: "Ready?... One, two three!" to him, then they both proceed to move the controls of the plane to create the effect of turbulence. Now, in the context of these activities, the *nonverbals* have provided us with the ingredients essential to make sense of the situation, and, in this case, to understand the humor behind it.

So what I'm really looking at is a particular subset of situations in which humans communicate in the context of actions; how I see the main issue can best be described in a Venn diagram (below).

Imagine two pilots operating an aircraft: each pilot interacts individually with the aircraft, via his or her own switches, controls, displays, and so on. At the same time, they also have certain situations where they interact with each other, exchanges of information which are almost independent of the machine. I am focusing my study on the place where they all overlap. If I were to replace my researcher's hat with my designer's hat, I would ask: "How will next year's cockpit better support the coordination between operators in that cockpit?"

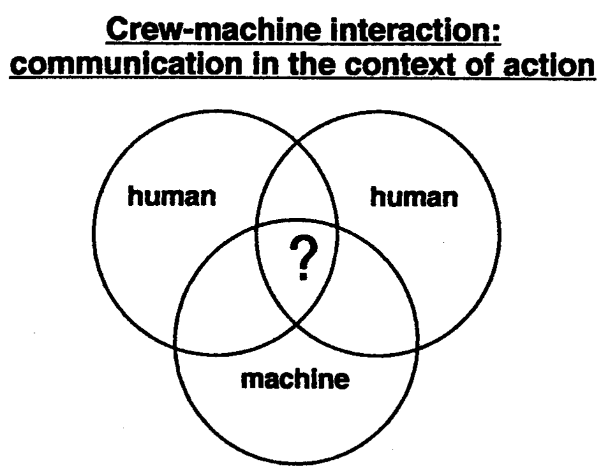


FIGURE 1: Crew-Machine Interaction:
Communication in the Context of Action

To take a broad view of this, I borrow a definition of "information" from Gregory Bateson, who says that: "Information is any difference that makes a difference." The question to be asked then is: "What differences make a difference within the context of flying an aircraft?" That information is what is shared between all crew members, and that is ultimately how they conduct their coordination, including the subset of information-exchange called speech communication. My position is that the reason speech communication works as well as it does is the richness and robust nature of information provided by the entire context.

Now, communication must be studied within a context—this is really the key, because information is always interpreted within a particular context. In the desert a tree is a landmark, in the forest it is not. It's a matter of context. When you're trying to find your way around the desert, you see that 1 tree out there, then see it on your map—that's it, you know exactly where you are. In the forest you may keep bumping into trees, and none of them tells you where you are. Similarly, the smell of fire means something different to firefighters. Again, it's a matter of context: if you're standing in the middle of a burning building, the smell of fire doesn't really inform you of much. If you're flying at night at 30,000 feet and you smell something burning... Well, you can see how there's a totally different meaning there.

Imagine you're driving a car—perhaps you're the person next to the driver, and you're navigating through a new city. There are different sources of information which you both share. In Figure 2, I describe all the different sources and categories of information that are present. I have divided the table into columns representing sources of information, and rows representing the different sensory modalities involved in perceiving that particular information.Let me walk you through the table. In the left-most column, I describe one source of information—the environment within which the car travels. This is one source of information to which both the driver and the navigator have access. Going down the rows we can look at the different categories of environmental information perceived by driver and passenger. There is visual information, such as signs, traffic, other cars, etc. There is auditory information, such as the bells at a railroad crossing or the siren of an ambulance. There is some kinesthetic information, such as the speed bumps before you get to a critical intersection, or feeling the slope of the road.

Finally, there are things you can smell: for example, there could be a brush fire near the road, so if you're driving through the smoke and your sense of smell says: "fire," since you've seen the source of the smell, there probably is no reason for alarm. You can see how there is an interaction between the different categories of information: seeing smoke resulted in a particular interpretation of the smell of fire. These lists are obviously not exhaustive; you can fill each cell with whatever other examples you like.

The second column deals with another source of information—the car itself. The fact that we're both driving in this particular artifact gives us some information. Obviously, the car has displays and switches which were designed to give you information whenever you look at them. Beyond the dedicated displays, you can see if the heater is on or off just by looking at the switch and observing it's position. There's also auditory information: you can hear the engine when

	Environment	**Car**	**Driver**
Visual	**Signs** **Traffic**	**Displays** **Switches**	**Pointing** **R.V. mirror**
Auditory	**RR Xing** **Sirens**	**Engine** **Blinkers**	**Speech** **Para-Ling.**
Kinesthetic	**Speed bumps** **Slopes**	**Accelerations** **Switches**	**Tap shoulder** **Correct input**
Olfactory	**Fire** **Orange blossoms**	**Fire** **Oil leaks**	**Alcohol**

FIGURE 2: What Makes Communication so Robust?
Information Redundancy

it revs up as it strains to go uphill; you can hear when the person has put their blinker on to turn right or left. What about kinesthetic information? Obviously you feel the accelerations. As drivers, you know how far to push a switch or to hit the small control because you do get some kinesthetic feedback that informs you of the proper position. You can smell important events: you can smell an electrical fire in the car, or you can check what type of oil leak you might have by distinguishing between the smell of hydraulic and engine oils.

Finally, there's a third source of information—the other person in the car, and that's where I come back to the issue of crew communication. As was the case in an aircraft cockpit, here, too, are 2 people interacting in a particular context; in the third column is the breakdown of available information; the different modalities that are involved in the exchange of information between these people. If you are sitting in the passenger seat and you see the driver look 3 or 4 times in rapid succession in the rear view mirror, it's either a police car or somebody who is perhaps driving too close. In this respect, seeing activity provides important information. Reaching for switches, pointing at landmarks along the road, straining to see a street sign—all of these activities, when observed, provide valuable, task-relevant information. From an auditory perspective, we can assume that driver and passenger are talking to each other, but let us not forget to include all the paralinguistics, including such things as: "mm," "uh," intonations, hesitations, and so on.

There are also kinesthetic cues: people tapping each other on the shoulder to call attention to something, or guiding a hand towards a particular switch. This happens more when you're teaching somebody to drive, but I know that in aviation it happens a lot when a student reaches for the wrong switch, and you see him or her going for the wrong thing, you just reach over and touch their hand and say, "No, that's not the right one." Finally, even smell plays a role, very often critical: you step into the car and you smell alcohol on your driver's breath.... You should definitely do something about that.

So what we have, just putting all of that together, is a whole environment—a particular task context—of redundant information within which speech, the auditory thing shared by the driver, is only one aspect. *All* of the information in this table serves for the interpretation of speech, and, obviously, speech can serve for the interpretation of other events. The interaction of these things is what I look at in the work I do. The rest of this talk will describe a particular study I carried out at NASA Ames Research Center; a simulator experiment which focused on the interaction between speech communication and activity information in an aircraft cockpit.

EXPERIMENTAL DESIGN

Now, I won't go into great details describing the experimental side. I have several papers that those of you who are interested are welcome to read (see publication list at the end of this paper). In general, this was a simulator study that was carried out at NASA Ames Research Center, in what's called the Advanced Concept Flight Simulator. What's important for the context of this particular talk is the experimental manipulation, which was the particular design of the interface of the checklist. As you know, most of what is performed in the cockpit is based on procedures that are detailed in a checklist; pilots must follow the checklist when they perform both normal and emergency procedures. The particular manner of interaction with any given checklist—the design of the interface between pilot and checklist—was the independent variable manipulated in this experiment. We had 2 primary types of checklist interface: the classical, hand-held, paper checklist, and a modern type of interface, in which the checklist was displayed on a touch sensitive CRT screen. A secondary distinction was made between 2 different types of touch sensitive screens: a manual and an automated version. In the manual condition, the interface required that the pilot perform every item on the checklist. In the automated condition, pilots shared the task with the system; the system performed some items, and the pilots performed others.

This is how the experimental procedure flowed: 24 pilots were put in a high-fidelity simulator and flown around in a very realistic scenario. All were experienced pilots, from the same airline; they were paired-off to form 12 crews. The crews were randomly assigned to 1 of the 3 conditions, thus creating 3 four-crew experimental groups: a paper checklist condition, a manual touch sensitive screen condition—in which they had to do every item—and an automated touch sensitive screen condition, in which they did some items and the machine, or the system, did some items for them. Figure 3 is designed to summarize the experimental design in a simple, "user-friendly," manner.

Now, before I go into the findings, let me describe the data we collected (this can be seen in Figure 4). We had 3 video cameras looking at the crews, recording all flights. One camera captured a wide view of both pilots, while the other 2 were directed and focused, respectively, at each of the 2 pilots. All the recorded video tapes were transcribed for both verbal and non-verbal activity. Additionally, we had an experienced flight instructor "fly" with all crews as an observer. Following each flight, the observer rated the crew's performance across various pre-defined per-

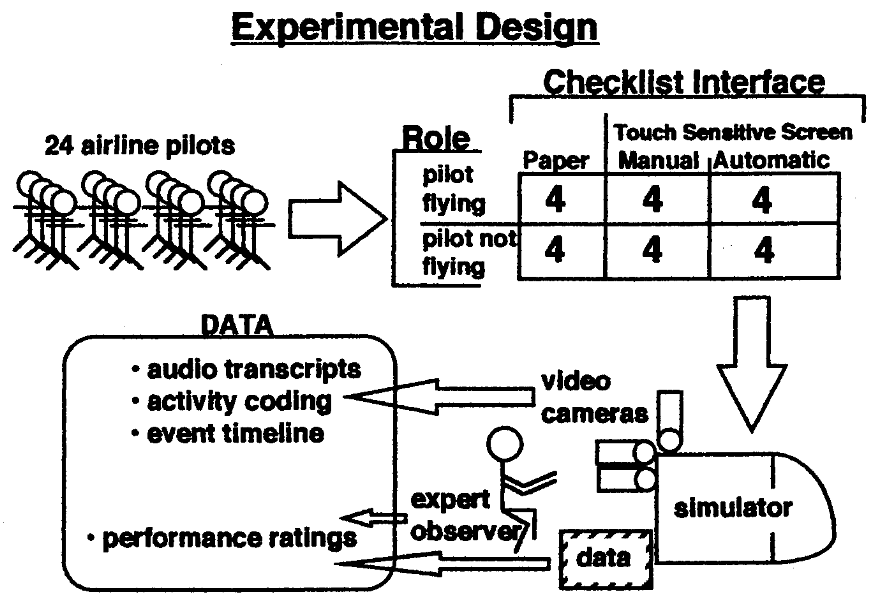


FIGURE 3: Experimental Design

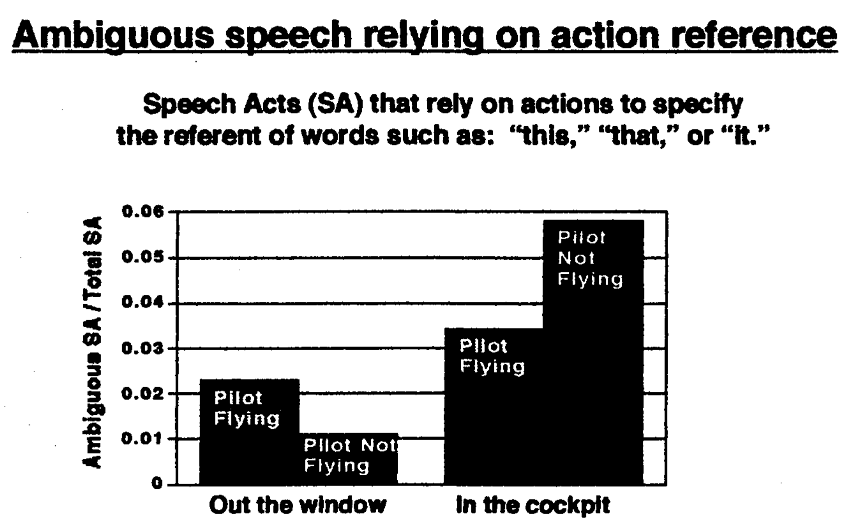


FIGURE 4: Ambiguous Speech Relying on Action Reference

formance categories and scales. As I go through the data, I will go into greater detail regarding what specifically was measured in every instance.

RESULTS

As part of the analysis, I first looked through the verbal transcripts for instances in which the words alone did not provide enough information which the listener could use to make sense of the utterance. In essence, I looked for utterances in which the words suggested that some form of non-verbal gesture or activity was being used to clarify utterances that were too ambiguous to stand on their own. For example, a speech act such as: "Watch out for this" may rely on the speaker's pointing to a particular dial in order to specify the referent for the word "this." Some other examples are: "Are you doing that?" or "What do you think of it?" These types of speech acts cannot be accurately interpreted unless one looks at the video and finds the activity that provided the key bit of information that gave that utterance operational meaning. What you see in Figure 4 is that the pilot flying the aircraft (PF) uses ambiguous statements in reference to objects outside the window more often than the pilot who is not flying (PNF). If you think of it, this makes sense. The task of flying the aircraft has the PF looking out of the window most of the time (this segment of the flight was performed in visual flight conditions), hence, he is usually the one to refer to objects in the environment. So when he says: "I hope we clear that mountain range," the PNF looks across the cockpit to see the pilot-flying's direction of gaze, thereby learning the particular range to which the speaker is referring.

Similarly, you can see that the PNF utters more ambiguous speech acts referring to items inside the cockpit. The reason for this is, I believe, that the PNF is so busy interacting with cockpit systems that for him or her, it is easier to just say: "This is wrong," than to explicitly specify a particular display or variable. For both pilots, the context of activity provides opportunities for communication using information that is inherent in their environment and actions.

Now, the question was, having found these results in the cockpit voice transcripts, will an analysis of nonverbal cockpit activity yield results that concur with these findings? In other words, do pilots actually use the shared context and capitalize on that particular form of visual information flow? To begin with, I decided to compare the rate of speech, as well as the rate of observed non-verbal activity, across all three groups (Figure 5). Overall, just looking at the effects of interface design on rate of speech, it seems that all 3 groups were virtually identical in the number of speech acts exchanged per minute. If you look at nonverbal activity, however, the rate of activity in the touch sensitive screen cockpit was significantly higher. This finding, in and of itself, does make sense. If you compare pilots sitting there going through a regular, hand-held, checklist with pilots who have to interact with a touch sensitive screen for each item they perform—in fact, not only perform the item but also to acknowledge to the system that the item has been performed—obviously, the latter group would be more active.

If you look at interactions with the checklist, and break them down to the activity performed by individual crewmembers (Figure 6), obviously, the PNF would be seen as more active than the PF. Note the clear difference between the paper group and the 2 touch-sensitive groups. Since the PFs in all 3 conditions were doing the same task—by virtue of the flight controls being identical in all conditions—one finds no difference between the 3 groups as far as these pilots are concerned.

Given that PNFs are visibly busier in the touch sensitive screen cockpits, is that increased activity used by their PFs to get added information? In that sense, do pilots monitor each other as they interact with the system? I looked at the total time that pilots just turned around to look at the other pilot's display, and then calculated the percent of that time that they turned while the other person was activating that screen (Figure 7). The PNFs, if you recall, were the ones who were the busiest interacting with the system. Accordingly, they almost never looked across the cockpit to see what the PF was doing, because there was no information there. There is not much to learn from looking over at a pilot who sits there with hands on stick and throttle. In contrast, the PFs were busy flying the plane, but they also had this extremely active PNF doing all these things on the checklist, and thus for them, it was very informative to turn and look at those actions. You can see that in the paper condition, there wasn't much to see, again, because the interface didn't afford too many visual clues. But with the touch sensitive interface, you see that over half of the time that the PFs turned to look at their crewmember was while that person was actually manipulating the checklist display.

I also used Penny Sanderson's MacSHAPA—which is a wonderful program—to look at the temporal connection between one pilot's activity and the other pilot's looking across the cockpit. My questions was: "How probable is it that when one pilot reaches for the display, the other pilot will turn around and look at that display?" Figure 8 describes this data, with expected probabilities on the horizontal axis, and the observed probabilities—based on the video transcripts—on the vertical axis. Note that it was significantly more probable than chance that a pilot will turn to look across the cockpit immediately following the other pilot's reaching for the checklist display.

Effect of design on speech and action:

Speech: Rate of speech for different interface cockpits is virtually identical.

Activity: Rate of activity in Touch Sensitive Screen cockpits is significantly higher.

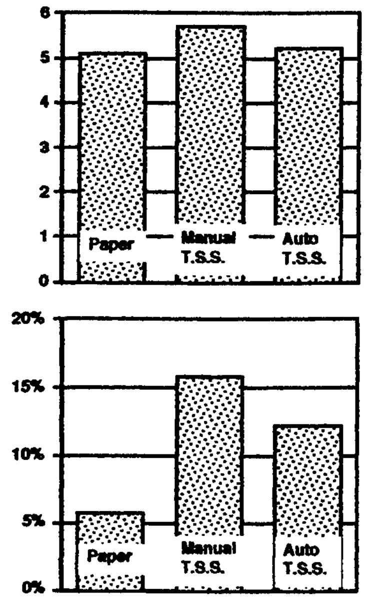


FIGURE 5: Effect of Design on Speech and Action

Distribution of workload between pilots

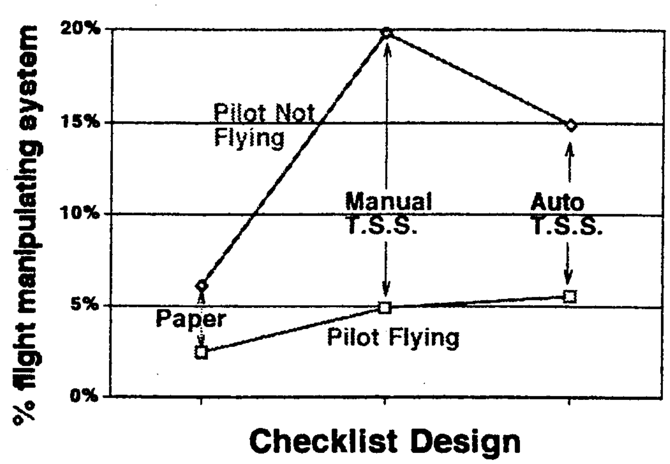


FIGURE 6: Distribution of Workload Between Pilots

Do pilots monitor other's activity?

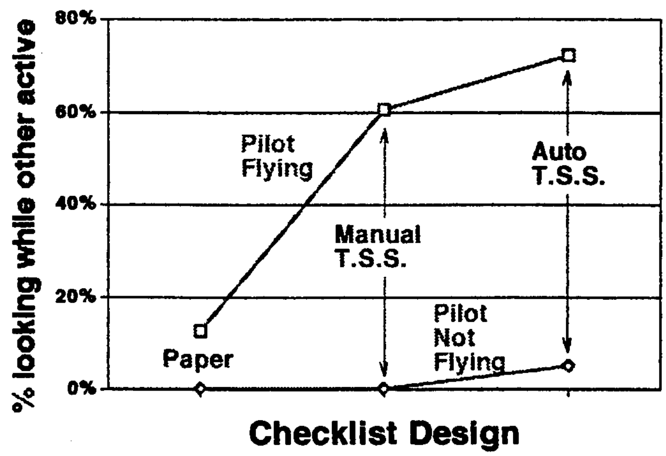


FIGURE 7: Do Pilots Monitor Other's Activity?

Does one pilot's action elicit a look from another?

In this study, pilots turned to look whenever their crewmembers reached to manipulate controls. As shown in this plot, this monitoring of crewmembers was performed consistently, at a significantly higher probability than expeted by chance glances across the cockpit.

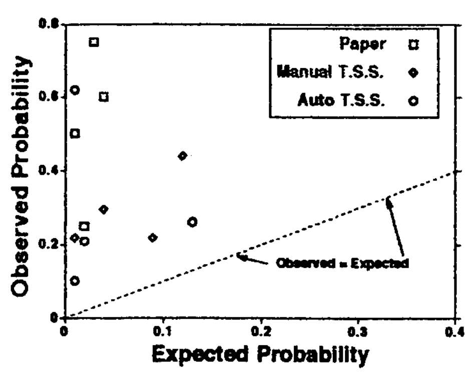


FIGURE 8: Does One Pilot's Action Elicit a Look from Another?

Performance ratings by observer

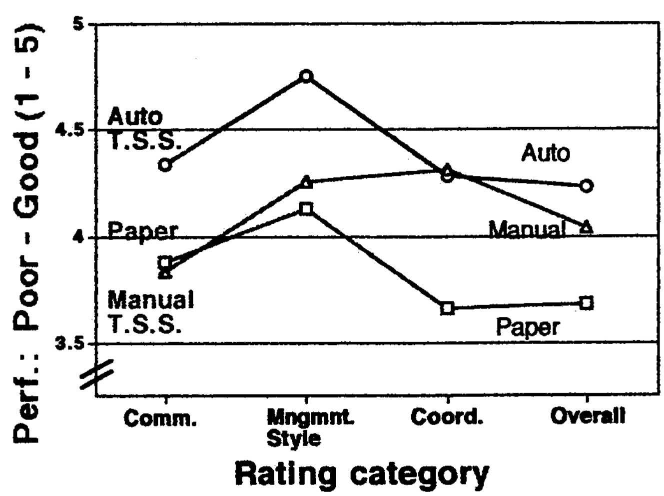


FIGURE 9: Performance Ratings by Observer

Cockpit design determines activity -- and resultant information

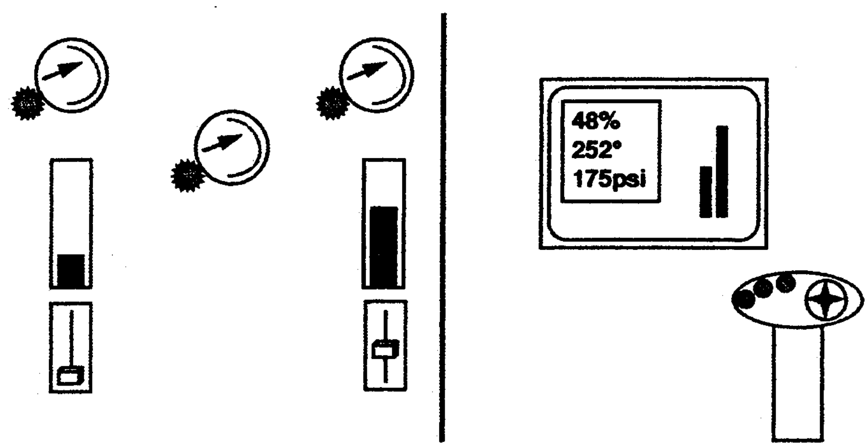


FIGURE 10: Cockpit Design Determines Activity -- and Resultant Information

We're actually seeing here an elicitation of a monitoring response by reaching for a particular display. So it seems pilots do indeed rely on visual information, turning to look at it whenever it is made available.

Finally, I'd like to briefly discuss performance. We had an in-flight observer sit and rate the crews for several different aspects of performance. These were what I call really "soft" categories of performance: communication, management style, coordination, and an overall score which attempted to capture whether the observer thought they were a good or a bad crew. What's interesting is that on these categories of communication, management style, and coordination, there were differences between the crews (Figure 9). This suggests that the style of communication was different to such an extent that the in-flight observer was sensitive to it, and thus produced data that suggests a systematic difference between the different conditions. The difference here between the automatic touch sensitive screen and the paper is significant; the manual group did not differ significantly from either one of the other 2.

CONCLUSIONS

So where do these findings lead us? I have an illustration that may help me clarify my perspective (Figure 10). As designers, we are at the point where we can design cockpits, for example, like the system on the left; whichever switch I reach for, you as a copilot can see precisely what I do. I don't have to tell you "I'm reaching for the switch that controls the up and down bar," because you can *see* it. If I fly with my hand on the stick and I have those variables controlled by my thumb—as is the case in the design on the right hand side—it's almost impossible to know what I'm controlling until the feedback has come back from the system to tell you that a change in the display *may* be a result of something I had just done. So as professionals who define and design the next generation of cockpits, we really want to decide what aspect of a task we want people to share, and perhaps, since some activity may be unnecessarily distracting, we need to decide what aspects of the task we do *not* want them to share. In that sense, we want to design critical information into the system, and also make sure we design redundant things out of the system.

To summarize, I'd like to call your attention to a figure I introduced earlier (Figure 1). The pilots are there in order to interact with this machine, controlling it according to their goals and the information and constraints provided by the environment within which that machine flies. There is a whole context, an environment of information, with signals and messages going back and forth between pilots and the environment, between pilots and the aircraft, and between the pilots themselves. When we look at speech communication, we are looking at a flow of information of a particular kind—a flow that takes place within the context of action and perception. Often in the cockpit, actions do speak louder than words. To better understand speech communication, we need to include in our scope other elements that affect the overall process of crew communication and coordination.

REFERENCES

Bateson, G. (1979). *Mind and Nature: A necessary unity.* E.P. Dutton, New York, NY.

Sanderson, P.M. (1993). *Designing for simplicity of inference in observational studies of process control: ESDA and MacSHAPA.* Proceedings of the Fourth European Conference on Cognitive Science Approaches to Process Control (CSPAC '93): Designing for Simplicity. August 25-27, Copenhagen, Denmark.

Segal, L. D. (1989). *Differences in cockpit communication.* In R.S. Jensen (Ed.), Proceedings of the Fifth International Symposium on Aviation Psychology. Columbus: Ohio State University.

Segal, L.D. (1993). *Automation design and crew coordination.* Proceedings of the 7th International Symposium on aviation Psychology. Columbus: Ohio State University.

Segal, L.D. (1994). *Effects of checklist Interface on Non-verbal crew communications.* NASA Contractor Report 177639. Moffett Field, California; NASA Ames Research Center.

* All actual accidents are reported to the NTSB and are not included in the ASRS database.

** It is not uncommon to receive accounts of the same incident from both a pilot and an air traffic controller. These multiple reports are included in the percentages.

INDEXICAL AND LINGUISTIC ATTRIBUTES IN SPEECH PERCEPTION: A REVIEW OF SOME RECENT FINDINGS

Ann R. Bradlow, Lynne C. Nygaard & David B. Pisoni
Speech Research Laboratory
Indiana University

INTRODUCTION

We begin this report with an introduction to the general approach and research topics that are currently under way in the Indiana University Speech Research Laboratory. After the introduction, we turn to a discussion of some recent work on the perceptual learning of voices, and on the relationship between voice attributes and speech intelligibility. This work addresses the following questions: What is it that you learn about a speaker's voice when you become familiar with that speaker's voice? What is it that you have acquired, or learned, about an individual's voice when you pick up the telephone and recognize the person at the other end of the phone? In the third part of this report, we will present some analyses of a large database of recorded sentences. Our interest in this study is the factors that influence speech intelligibility. In this study, we ask questions such as the following: What makes one talker more intelligible than another? And, what makes one sentence easier to recognize than another? Finally, we end with some general concluding remarks about sources of variability in speech perception and production.

We are currently working on several projects in the Indiana University Speech Research Laboratory. One of the major interests of our research deals with spoken word recognition and the mental lexicon, which we think of as the interface between the sensory input and comprehension. We're particularly interested in issues that have to do with variability in speech and how it influences word recognition and speech perception. We've also become interested in issues about perceptual learning and adaptation, particularly adaptation to voices, to changes in speaking rate, and to other aspects of individual talkers that modify the way they talk. Finally, we have an ongoing interest in developing new techniques for studying online, or real-time, comprehension of spoken language. Thus, the general kinds of problems that we're interested in deal with the nature of lexical knowledge, and the neural representation of speech in memory.

In general, we are concerned with the physical properties of spoken language, which can be approached by studying speech in 3 interlocking domains. The first of these domains is the articulatory aspect of speech, that is, the ways in which people physically produce speech. Studies of articulation generally use various kinds of physiological measurement techniques. The second domain of speech is the acoustic domain, that is, the domain of the acoustic consequences of speech articulation. These studies generally involve acoustical measurements of the speech signal. Thirdly, one cannot study speech without also studying its consequences in the perceptual domain. These second and third domains, the perceptual and acoustic domains, have been the focus of our main research interests.

Finally, in this introduction, we give an overview of the general theoretical framework within which we conduct speech research in our laboratory. Most of what we know about speech and language has been approached from an abstractionist, or symbolic, orientation, which has been motivated primarily by the transformational approach to linguistics. This formal approach to language views much of the personal, or "indexical", properties of speech as irrelevant to the neural processing of speech signals by the auditory system. Examples of these "indexical" characteristics are gender, dialect, speaking rate, physical states, emotional states, age, weight, and social status. Morris Halle (1985) voices this position in the following quotation:

> When we learn a new word, we practically never remember most of the salient acoustic properties that must have been present in the signal that struck our ears. For example, we do not remember the voice quality of the person who taught us the word or the rate at which the word was pronounced, not only voice quality, speed of utterance and other properties directly linked to the unique circumstances surrounding every utterance are discarded in the course of learning a new word.

In contrast to this approach, we believe that these talker- and instance-specific characteristics are all intimately intertwined in the acoustic signal, and that they are involved in the perceptual analysis, encoding and storage in memory of the speech signal. The studies we present below provide support for this position.

Perceptual Learning of Voices

In this section, we present a summary of recent perceptual experiments that address the issue of how listeners contend with variability in the speech signal, in particular, variability due to talker characteristics. The traditional approach has been to think of this as a "normalization process," that is, as a process that involves a "stripping away" of the acoustic variability in the signal to arrive at a set of canonical, idealized, symbolic linguistic units. This approach to speech assumes that the variation due to talker characteristics is discarded in developing a representation of the speech signal. In contrast, our approach views this acoustic variability as an important source of information for the listener that is not lost, but rather incorporated in a long-term representation of the talker's utterance.

An explicit description of this dichotomy was introduced by Laver (1989) and Laver and Trudgill (1979) who contrasted "linguistic" and "indexical" factors. The "linguistic" factors of an utterance are characterized by the formal, symbolic units that are hypothesized by the listener. This linguistic content of an utterance serves a communicative purpose in that it conveys the message intended by the sender to make the receiver aware of something. The "indexical" factors of an utterance convey information such as the identity and attitudinal state of the speaker. These factors serve to convey information about the speaker regardless of the intentions of the sender.

Our goal in these perceptual studies was to investigate the relationship between the processing of talker information and the processing of the linguistic content of a speaker's utterance. In particular, we wanted to know whether familiarity with the talker's voice would affect the processing of words spoken by that talker. Using isolated words spoken by 10 talkers, we trained listeners to recognize the talkers' voices (see Nygaard et al., 1994). It took about 9 days to get a group of subjects up to a criterion level of 70 percent correct talker identification. At the end of the training period, we investigated the perception of spoken words by asking the listeners to recognize the words rather than identify the voice characteristics of the talkers. We hypothesized that familiarity with the talker's voice would affect subsequent word recognition, and in so doing, would provide evidence for a direct link in processing between encoding of talker information and spoken word recognition. Note that in this experiment we had 2 conditions. Subjects in the first condition were trained to identify a set of voices and then performed the word identification task with the same set of voices (the now familiar voices). Subjects in the second condition were trained to identify a set of voices and then tested in the word identification task with a set of unfamiliar, or novel voices. To assess the effect of talker familiarity on word recognition, we compared the performance on the word identification task across the 2 groups.

Before we discuss the results of this experiment, we need to review in more detail the specifics of the experimental procedure. During the 9-day training period, listeners were trained to recognize each talker's voice and to associate that voice with 1 of 10 common names. There were 3 phases to this training period. First, the subjects just listened to the voices and tried to remember the names of the talkers. Next, the subjects performed a voice recognition task with feedback. Finally, in the third phase of the training period, subjects performed the voice recognition task without feedback. On the tenth day of the experiment, subjects were given a generalization test. This test assessed whether the knowledge the subjects had obtained from the talker's voice during training was specific to the words used in training. Thus, the stimuli for this test of generalization were novel words (i.e. words not used in the training period) produced by the same 10 talkers used in training, and subjects were asked to identify the talkers' voices. Subjects received no feedback in this test. After the test of generalization, subjects performed the word intelligibility test. This final test was the crucial test for determining whether the ability to identifying the voice transfers to a completely different type of task, that is, to identifying the linguistic content of what the talker was saying. This word intelligibility test presented the listeners with novel words; and, they were asked to identify the word rather than the voice. The words were presented at 4 different signal-to-noise ratios.

Figure 1 shows the time course of the subjects' performance from the start of the training period to the test of generalization. Data from the 2 groups of subjects are shown separately. Both groups were trained on the same set of voices. The "trained" group was then tested on the familiar voices in the word intelligibility test; whereas, the "control" group performed the word intelligibility test on a different set of unfamiliar voices.

As shown in Figure 1, this is a very difficult task for listeners. Assuming that listeners are able to distinguish speakers on the basis of gender right away, the chance level of performance for the 10 talkers is 20 percent. Thus, at the start of training, subjects are a little above chance. Both groups then learned to identify the voices at about the same rate over the 9 days, and they then generalized quite well to novel words produced by the same talkers. The data shown in this figure are only for those subjects who reached a set performance criterion of 70 percent correct on the ninth day of training. Our reasoning behind setting this criterion was that we couldn't assess the subjects' transfer from the voice identification task to the word

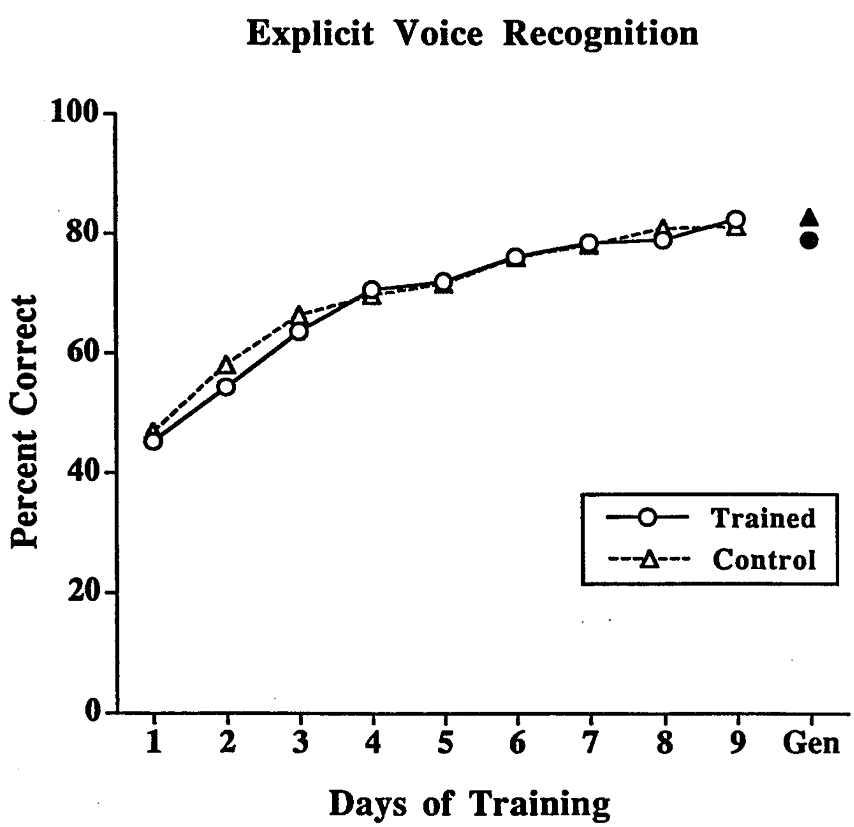


FIGURE 1: Training on Explicit Voice Recognition

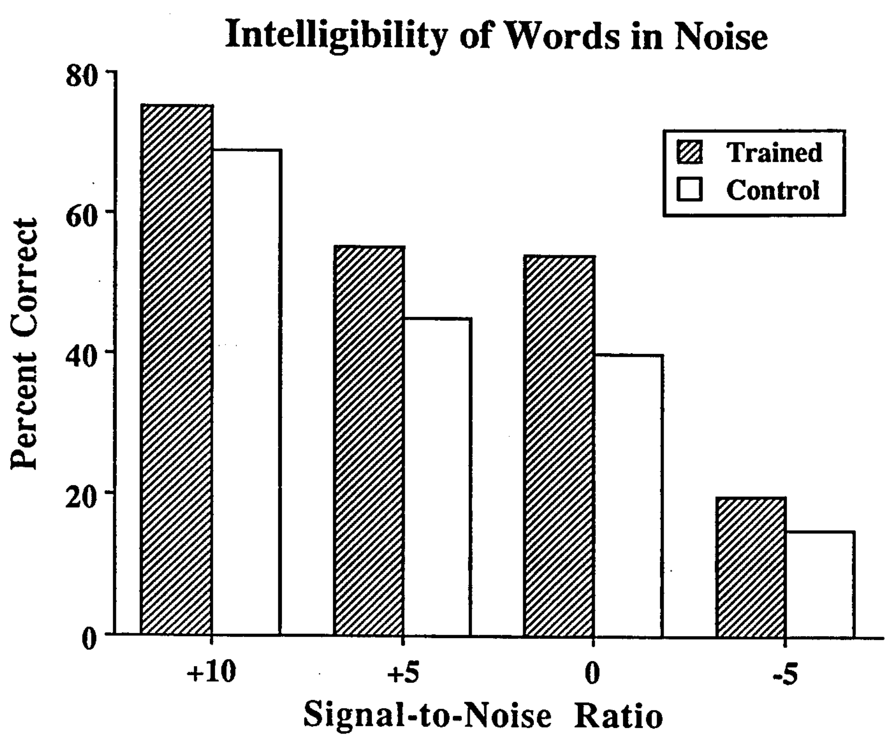


FIGURE 2: Transfer of Training on Voice Recognition to Word Identification in Noise

intelligibility task if they hadn't actually learned the voices. Thus, the data shown in Figure 1 are only from the subjects that reached this level of voice identification. These were approximately half of the original set of subjects.

Figure 2 shows the data for the word intelligibility test, in which subjects were asked to transcribe a set of novel words. In this task they were required to attend to the linguistic content of the word, rather than to the voice of the talker saying the word. The figure shows the accuracy data at each of the 4 signal-to-noise ratios for both the "Trained" and the "Control" subject groups.

At each signal-to-noise ratio, we found that the group of "Trained" subjects (those who identified words spoken by familiar voices), performed better in this transfer task than the group of "Control" subjects (those who identified words spoken by unfamiliar voices). This result demonstrates that people are better able to identify words that are produced by talkers that are familiar to them, and this suggests that voice recognition and the processing of the phonetic content of a linguistic utterance are not independent. The implication of this result is that experience with specific acoustic attributes of a talker's voice significantly improves spoken word recognition.

Given these results, we now ask what kind of knowledge is acquired when listeners are learning to recognize voices? In response to this question, we put forward three proposals. (There are, of course, several others that might merit consideration.) First, we consider the possibility that in learning to recognize voices, listeners are acquiring a form of procedural knowledge (Kolers, 1976; Kolers & Roediger, 1984). Within a framework that assumes a normalization process to handle talker-specific variation, this proposal suggests that listeners learn (and retain in memory) the normalization process that is applied to a talker's voice. Listeners learn to "unravel" the talker-specific information from the linguistically meaningful information, and this learning of specific perceptual operations that compensate for talker-specific variation facilitates further processing. A second proposal is that the listeners learn specific sets of features or attributes of the talker's voice and that these attributes are encoded in memory. Characteristics such as fundamental frequency, relative formant spacing, and glottal attributes may be stored in a memory representation for a talker's voice and used as a reference or template for subsequent phonetic processing. Finally, a third proposal is that listeners learn something more abstract. Listeners may become sensitive to information in the acoustic signal about specific dynamic properties of the talker's vocal tract as an acoustic source (Remez et al., 1981).

The results of this experiment also have some important implications for current theories of speech perception and spoken word recognition. First, these results suggest that representations of spoken words in memory may be much more detailed than previously thought. Second, any proposed mechanism of perceptual compensation in speech must be susceptible to general processes of perceptual learning and attention. Finally, explanations of speech perception and spoken word recognition may need to include the role of long-term memory for source characteristics.

In a follow-up set of experiments that are currently under way in our lab, we are investigating the specific type of training that leads to the talker-familiarity advantage that we obtained in the experiment reported here. We are interested in seeing if it is mere exposure to a talker's voice that facilitates word identification, or if listeners must explicitly attend to voice attributes during learning to facilitate linguistic processing. We are using a word identification training task that can then be compared to the voice learning training task.

Instance-specific correlates of speech intelligibility

In this section, we present a study that is currently under way to determine some of the instance-specific correlates of speech intelligibility. We are working with a multi-talker sentence database that includes recordings of 100 Harvard sentences produced by 20 talkers (10 males and 10 females), giving a total of 2000 recorded sentences. The sentences all consist of 1 main clause with 5 keywords and a variable number of "filler" words in between these 5 keywords. Along with this production data, the database includes intelligibility scores for each talker's production of each sentence. This intelligibility data was collected by having 10 listeners transcribe each talker's production of each of the 100 sentences. Thus, we had 10 listeners per talker, giving a total of 200 listeners. The transcription data was scored using a criterion that counted a sentence as correctly transcribed, if and only if, each of the 5 key words was correctly transcribed. All other sentences were counted as incorrect. This data provided us with a means of exploring some of the sources of variability in sentence and talker intelligibility.

Figure 3 shows the variability in sentence intelligibility across the 100 sentences. The sentence intelligibility scores shown in this figure are averaged across all 20 talkers and all 10 listeners per talker. It is clear from this plot that there is considerable variability in overall sentence intelligibility.

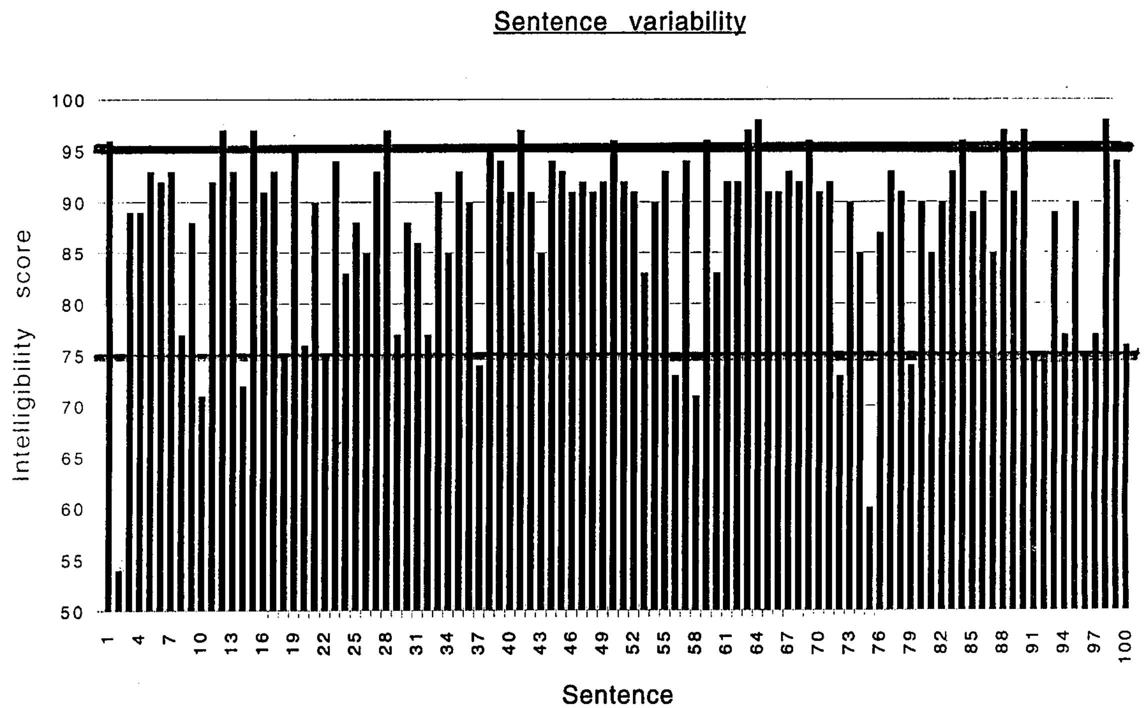


FIGURE 3: Variability in Sentence Intelligibility

The question we posed here is: What makes one sentence more intelligible than another? In order to address this question, our strategy was to compare a set of high-intelligibility sentences with a set of low-intelligibility sentences. The high-intelligibility sentences were all of the sentences that had overall intelligibility scores above 95 percent (above the upper line in Figure 3), and the low-intelligibility sentences were all of the sentences that had overall intelligibility scores below 75 percent (below the lower line in Figure 3). This gave a set of 14 high-intelligibility sentences and a set of 9 low-intelligibility sentences that we could compare in terms of sentence length and various other lexical characteristics of the component keywords.

The first result of this comparison is that on average the high-intelligibility sentences have fewer words than the low intelligibility sentences (7.21 versus 8.22 words per sentence). Since the scoring criterion is based on the correct transcription of the keywords, this result implies that keywords that are embedded in longer sentences are more susceptible to transcription error than keywords that are embedded in shorter sentences. A second difference between the words in the high-intelligibility sentences and the words in the low-intelligibility sentences is the number of "function" versus "content" words as sentence keywords. A function word is a closed-class word that is morphologically simplex, such as pronouns, prepositions, and articles; content words are open-class words that can be morphologically complex, such as nouns, verbs and adjectives. In our multi-talker sentence database, we found that the high-intelligibility sentences have more function words as keywords than the low-intelligibility sentences (21% versus 11%). A consequence of this difference in the lexical status of the keywords across the 2 sets of sentences is that the keywords in the high-intelligibility sentences have a higher mean frequency (1063.73 versus 152.31 occurrences per million words of printed text) and are shorter (3.6 versus 4.1 phonemes per word) than the keywords in the low-intelligibility sentences. (Function words are generally more frequent in the language and shorter in length than content words.) So far, we've seen that both sentence length and the type of words that comprise the sentence may contribute to making it a high- or a low-intelligibility sentence.

Another sentence-related attribute that we looked at in the comparison between the high- and low-intelligibility sentences has to do with the "neighborhood" characteristics of the words. As shown in the schematic in Figure 4, the "similarity neighborhood" of a word is defined as the set of words that differ from the target word by a 1 phoneme deletion, substitution, or insertion. For example, the word "can" has as neighbors "ban" (by a 1 phoneme substitution), "an" (by a 1 phoneme deletion), and "scan" (by a 1

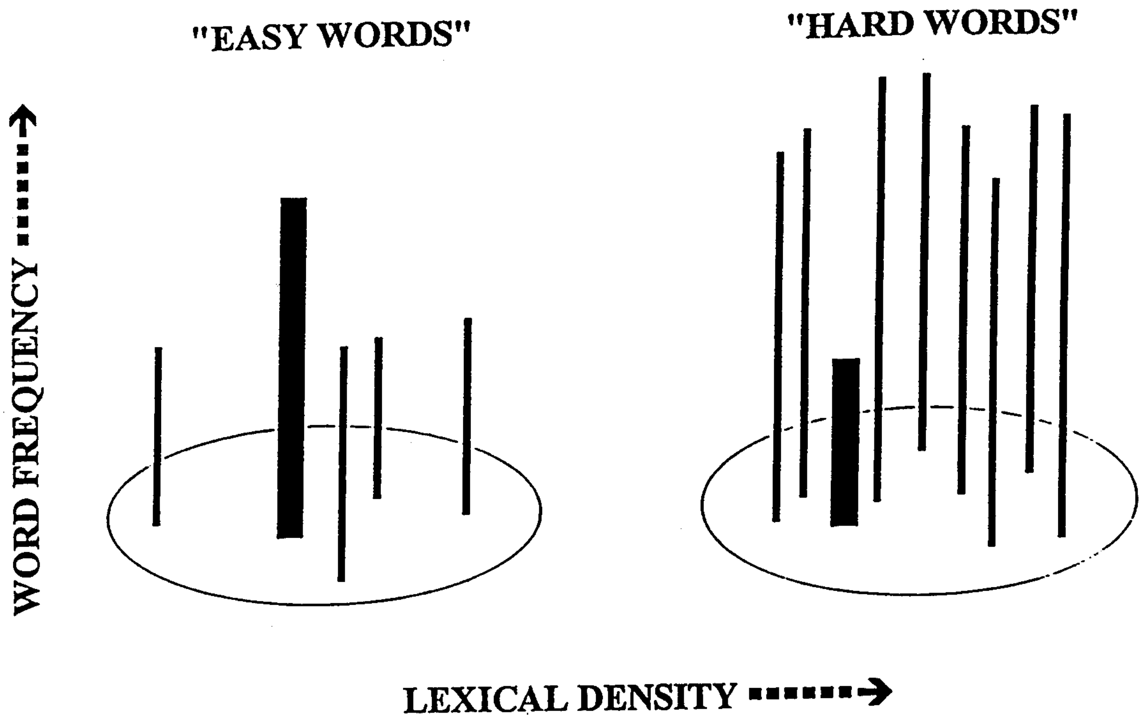


FIGURE 4: Schematic Representation of Lexical Similarity Neighborhoods

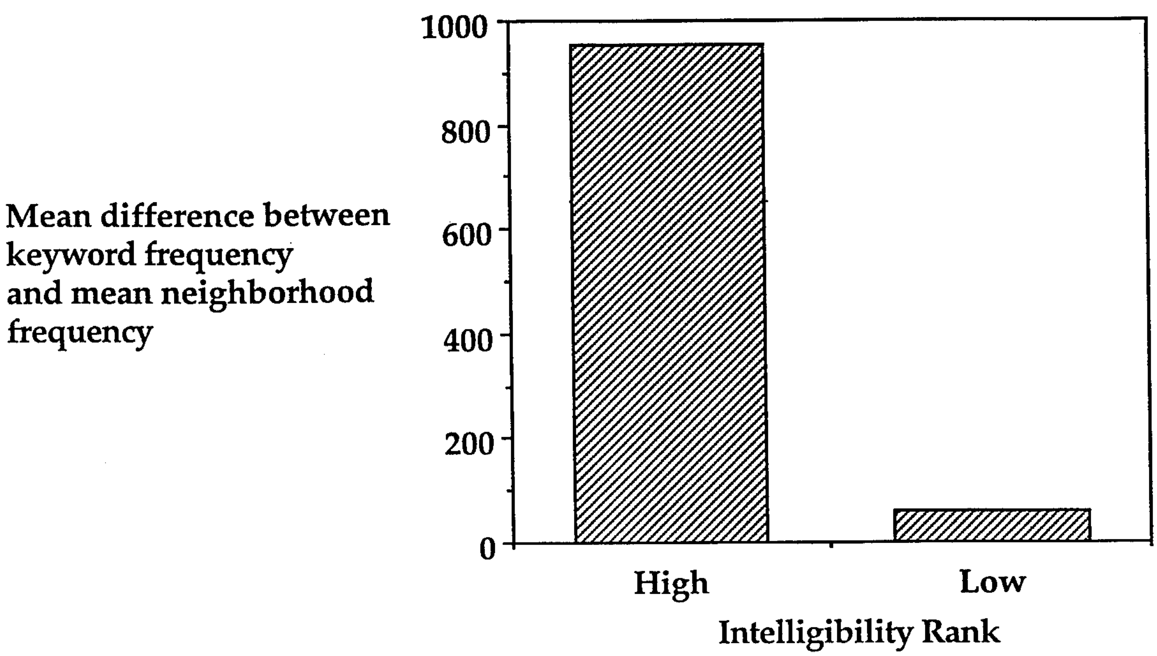


FIGURE 5: Mean Difference Between Keyword Frequency and Mean Neighborhood Frequency for Words in High- and Low-Intelligibility Sentences

phoneme addition). In other words, the similarity neighborhood of a word is the set of words that are phonetically similar to the target word. In Figure 4, the circles show the bounds of the similarity neighborhoods of 2 target words (represented by the thick bars). The vertical axis in this figure represents word frequency. In the case of the "easy" words, the similarity neighborhoods are sparsely populated (there are few phonetically similar words), and the frequency of the target word is considerably higher than the frequency of its neighbors. In other words, "easy" words "stick out" and are "prominent" in the neighborhood. In contrast, "difficult" words have many neighbors - they come from a densely populated neighborhood - and the target word frequency is considerably lower than the mean frequency of its neighbors. This "hard" word is "swamped" by its neighbors.

Previous work has shown that these neighborhood characteristics affect recognition of isolated words (e.g. Luce, Pisoni, and Goldinger, 1990). Thus, we wondered whether these characteristics of words embedded in sentences would effect the overall sentence intelligibility. In fact, in our multi-talker sentence database we found that the mean difference between keyword frequency and neighborhood frequency is greater for the words in the high-intelligibility sentences than for those in the low-intelligibility sentences (Figure 5). In terms of density, the neighborhoods for the words in the 2 sets of sentences are about the same. Thus, the words in the low-intelligibility sentences are "harder" than those in the high-intelligibility sentences.

We now turn to a discussion of the variability in talker intelligibility. Recall that in this database we had 20 talkers (10 males and 10 females) produce the same set of 100 sentences. Figure 6 shows the overall intelligibility for each talker averaged across all 100 sentences and all 10 listeners per talker. As shown in this figure, there is considerable variation in overall intelligibility across talkers; some talkers are generally more intelligible than others. Our question here is, "What are some of the characteristics of these talkers that make one talker more intelligible than another?" The talker characteristics that we compared across talkers in our database are gender, overall speech rate, and some details of phonetic timing.

We begin by investigating gender-related differences in talker intelligibility. The motivation for looking at gender as a talker characteristic that may play a role in overall intelligibility comes from a claim in the literature that female speakers exhibit fewer phonological reduction phenomena than male speakers (Byrd, 1992). The prevalence of reduction phenomena, such as increased speaking rate, unreleased final stops, alveolar stop flapping, and unstressed vowel reduction, is generally associated with a less formal, conversational, even "sloppy" style of speech. Thus, we wondered whether in our database we could find a gender-related difference in intelligibility which might be related to gender-based difference in the prevalence of reduction phenomena. In fact, we found that our female talkers generally had higher intelligibility scores than the males (89.4% versus 86.3%, $p(18)=0.02$). Furthermore, all 3 of the highest intelligibility talkers were female, and all 3 of the lowest intelligibility talkers were male.

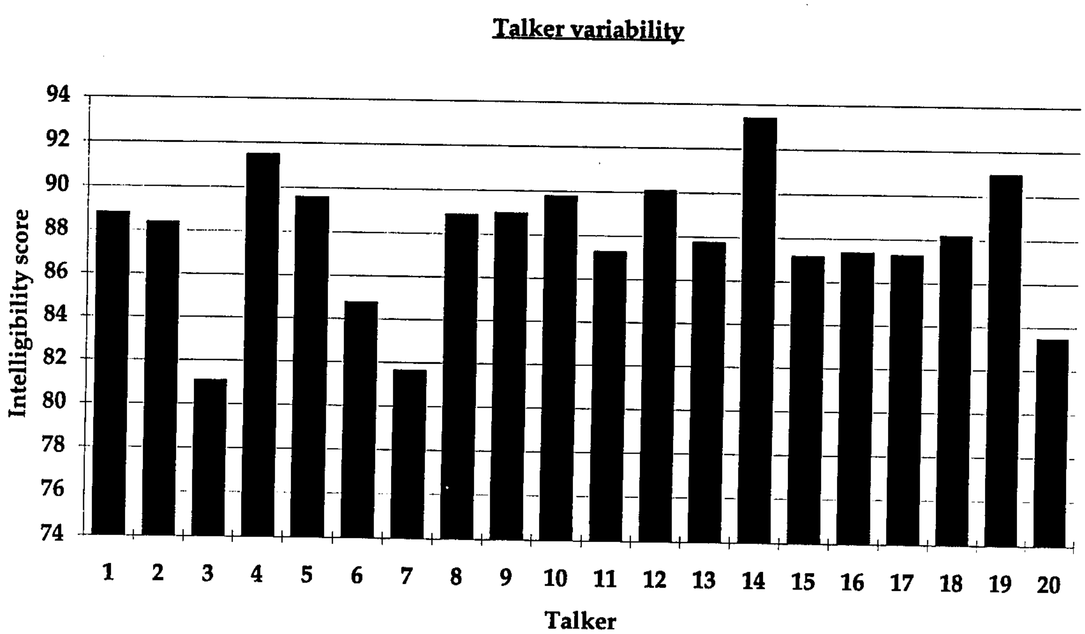


FIGURE 6: Variability in Talker Intelligibility

In order to investigate the link between the prevalence of reduction phenomena and talker intelligibility, we began by asking whether overall speaking rate correlates with talker intelligibility. In our database, we found that the mean sentence duration for the 3 high-intelligibility talkers is indeed greater than the mean sentence duration for the 3 low-intelligibility talkers. In other words, we found that the best and the worst talkers differ in overall speaking rate. However, we also found that the 10 males had longer sentence durations than the 10 females in spite of the fact that the females generally had higher intelligibility scores than the males. Furthermore, across all 20 talkers, there was no correlation between overall speaking rate and intelligibility. Thus, it appears that differences in overall speaking rate do not correlate very well with differences in overall intelligibility. We do find that at the edges of the distribution of intelligibility scores, overall speaking rate is a differentiating factor (the talkers with the highest intelligibility scores do have slower overall speaking rates than the talkers with the lowest intelligibility scores). However, if we consider the full set of 20 talkers, the rate-intelligibility correlation does not hold. This result led us to begin investigating some of the finer details of phonetic timing in the speech signal to see if temporal factors at a more detailed level might correlate with overall talker intelligibility.

In order to investigate the fine-grained details of phonetic timing, our approach was to investigate cases of consistent listener error. For example, a common listener error in the phrase "the walled town" involved simply failing to detect the medial /d/. Many listeners transcribed this phrase as "the wall town." Our question here is, "What factors in the talker's productions of this phrase determine /d/ detection?" In order to address this question, we measured various portions of the time dimension for the period of the signal that covers the /dt/ sequence in this phrase. This period extends from the onset of the closure for the /d/ to the offset of the aspiration for the syllable initial /t/. In almost all cases, the talkers produced 1 stop with a /d/-like closure and a /t/-like release, rather than releasing the /d/ before forming a second closure for the /t/.

We begin by looking at the correlation between the total duration of this vowel-to-vowel period (the /dt/ sequence) and the rate of /d/ detection by the listeners for each talker. This correlation is positive (Spearman rho = 0.702). In order to investigate what part of this /dt/ sequence correlates best with /d/ detection, we examined the /d/ closure and /t/ release portions separately. The analysis showed that it is the /d/ closure portion that correlates with the rate of /d/ detection (Spearman rho = 0.768 versus 0.211 for the /t/ release portion). Finally, in order to see what part of the /d/ closure correlates with rate of /d/ detection, we examined the voiced and silent portions of the closure separately. Here we found that it is the absolute duration of the voicing during closure that correlates with the rate of /d/ detection for each talker (Spearman rho = 0.744 for the voiced portion versus 0.225 for the silent portion). We also examined the correlation between rate of /d/ detection and the voiced closure duration relative to the total closure duration, relative to the entire /dt/ duration, and relative to the durations of the preceding and following syllables. However, none of these proportional measures correlated with rate of /d/ detection. Thus, the longer the voiced closure in an absolute sense, the more likely it is that listeners will detect the voiced consonant in the /dt/ sequence. This case presents an example of variation across talkers at the level of phonetic implementation that has an important effect on the talker's intelligibility.

Another case that we looked at in this manner was for the phrase "the play seems" which was often transcribed as "the place seems." In this case, the error was in determining the syllable affiliation of the /s/. We measured the duration of the /s/ and correlated this duration with the rate of correct /s/ syllabification. The results showed that the absolute duration of the /s/ had a small negative correlation with the rate of correct syllabification (-0.254); however, this correlation strengthened when the /s/ duration was taken as a proportion of the preceding word duration (-0.653). Thus, it appears that the longer the /s/ relative to the duration of "play," the more likely it is to be incorrectly syllabified as both the final consonant of the preceding word and the initial consonant of the following word, giving "place seems" rather than "play seems." In this case, the talker needs to be very precise in the timing relation for the listener to correctly interpret the signal. Additionally, in this case the more carefully articulated form is the shorter form, thus providing a possible explanation for the poor correlation between slower overall speaking rate and higher intelligibility scores. Furthermore, in this case we found that the female talkers were generally more accurate in executing this timing relation than the male speakers, and consequently there were fewer transcription errors for the females than for the males.

The general finding of this exploratory study suggests that listeners are indeed sensitive to the fine-grained details of an utterance to the extent that they affect overall intelligibility. Both sentence- and talker-related characteristics are detected by the listener and interact with the processes of speech perception to affect speech intelligibility, rather than being separated from the signal at an early stage of phonetic processing.

CONCLUSION

In these concluding remarks we'd like to stress the approach that we have taken in our laboratory with regard to the role of variability in speech perception. It is probably fair to say that the attitude of many speech scientists since the beginning of modern speech research after World War II, was that acoustic variability in the signal is not informative to the listener. Typically, variability has been thought of as something that needs to be eliminated by the processes of speech perception. Thus, many of the now classic experiments in speech perception were designed with stimuli from a single talker who read a list of words in citation form in a benign recording environment. The results of such studies were therefore very difficult to generalize to real world environments where listeners are operating in very robust conditions. Similarly, the reliance on highly simplified, idealized, synthetic stimuli has given us a very misleading understanding of the way human listeners operate in highly variable environments. In fact, synthetic stimuli are very impoverished signals, and human listeners have evolved over the ages to deal with very robust and highly redundant signals. In contrast to this traditional approach, we believe that variability is informative and is an aspect of the signal that is not only *not* eliminated by the listener, but is actually encoded as part of the neural representation of speech. The studies that we reported here demonstrate that this information is encoded and used by listeners in a variety of behavioral tasks. Furthermore, this work presents examples of studies that were designed within a non-symbolic, non-abstractionist theoretical framework that focuses specifically on how human listeners cope with variability in the perceptual environment. Rather than designing experiments which eliminate variability, as done in the past, our approach to research has been to design experiments which specifically incorporate a very substantial amount of variability. In some related clinical work that we are currently pursuing at the Indiana University Medical Center in Indianapolis, we are developing a new battery of tests for hearing impaired listeners which we call PRT tests (perceptually robust tests). The idea here is to study clinical populations with tests that are designed to imitate real-world conditions where there is an enormous amount of variability due to factors such as speaking rate, ambient noise, and different voices (Kirk et al., in press).

The work we described in this paper provides support for a new approach to perception and cognition, which is associated with Larry Jacoby and Lee Brooks and a number of other theorists. This non-analytic approach to cognition is based on the idea that we store very fine details in memory and that we use specific instances rather than engaging in processes of abstraction. As suggested by the data presented above, we believe that much of the phonetic and highly variable acoustic-phonetic detail that is present in the speech signal is, in fact, encoded in memory and used in the process of speech perception and spoken language processing. We believe these findings on the role of variability in speech perception have a number of important implications for research, theory and clinical applications.

REFERENCES

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (In Press). On the contribution of instance-specific characteristics to speech perception. In J. Schoentgen, J.M. Ramlot, C. Sorin, H. Méloni, & J. Mariani (Eds.), *Levels of Speech Communication: Relations and Interactions.* Amsterdam, Netherlands: Elsevier.

Byrd, D. (1992) *Sex, dialects, and reduction.* In J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodge, & G. E. Wiebe (Eds.), Proceedings 1992 International Conference on Spoken Language Processing. Alberta, Canada: University of Alberta.

Halle, M. (1985). Speculations about the representation of words in memory. In V. A. Fromkin (Ed.), *Phonetic Linguistics.* New York: Academic Press.

Kirk, K. I., Pisoni, D. B., & Osberger, M. J. (In press). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear and Hearing.*

Kolers, P. A. (1976). Pattern analyzing memory. *Science*, 191, 1280-1281.

Kolers, P. A. & Roediger, H. L. III. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior,* 23, 425-449.

Laver, J. & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K. R. Scherer and H. Giles (Eds.), *Social Markers in Speech.* Cambridge University Press.

Laver, J. (1989). Cognitive Science and speech: a framework for research. In H. Schnelle & N.-O. Bernsen (Eds.). *Logic and Linguistics: Research directions in Cognitive Science. European Perspectives, Volume 2.* Hillsdale NJ, Erlbaum Associates.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. Altmann (Ed.). *Cognitive Models of Speech Processing: Psycholinguistics and Computational Perspectives.* Cambridge , MA: MIT Press.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5, 42-46.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, 212, 947-950.

SPEECH RECOGNITION IN THE TOWER CABS AND WHAT ALLOWS IT TO WORK

Martin Thee
Air Traffic Control System Support Branch
FAA Academy
Federal Aviation Administration

THE SYSTEM

When the tower cab simulators were first installed at the Mike Monroney Aeronautical Center, the confidence level of the instructors was tentative at best. Instructors who expected excellent recognition results experienced a few problems, the majority of which were blamed on the voice recognition system. Most trouble calls that we received indicated that a particular student (and in some cases the entire lab) was getting 0% recognition. A statement of the problem was generally followed by the question: "*What is wrong with the system?*" We suggested that the instructor adjust the mike for the student (or a new noise calibration for the entire lab) and that frequently took care of the problem.

THE USER

After several months of successful resolution of trouble calls, instructor attitudes began to change. The calls were coming in less frequently and the question changed to "*Could you come and watch this student and see what is causing mis-recognition?*"

What changed their attitude

As instructors became aware of potential voice recognition problems and how to avoid them, their students consistently got good recognition. Logically, since instructors were able to get good recognition consistently (without ever re-training words) students should be able to get good recognition also. Instructors realized that poor recognition could be corrected with minor adjustments or some words of advice to the student.

PREVENTING RECOGNITION ERRORS

Prep before Enrollment

Before the enrollment process begins, all students are given a *30-minute briefing* on how the enrollment process works. Each student reads off some example digit scripts, and some example enrollment scripts and phrase scripts.

Digit Training

Students read 2 or 3 digit strings with an *experienced* instructor who listens for correct pace, volume and authority, before keying the mike and beginning digit training. Students are watched (i.e., listened to) closely during the digit training phase. If the student is having any problems during this phase (i.e., it is taking more than 10 minutes), the student is re-started at the beginning of digit training. For example:

"ONE TWO FOUR THREE FIVE"..."NINE SIX EIGHT ZERO SEVEN"

Carrier Word Enrollment

Students are instructed to speak each carrier word phrase as if it were a complete sentence. Instructors stress that the student should not pause between words, the words should flow smoothly.

Enrollment Process

Students are still being watched closely during the first few minutes of the Enrollment process. *Students are reminded* that if they pause between words, or "*bounce*" the words they may have to repeat the entire 2-hour enroll/train process.

Phrase Training

Phrase training (also known as "in context" training is the final stage of the training process. By the time the students enter phrase training they are on their own. In the briefing they are told to speak each phrase as if they were in a control tower and *directing* aircraft motion (as opposed to requesting aircraft motion).

Mike Check

When anyone sits down to use voice recognition they must perform a *Mike Check*. During a mike check, the mike is turned on, the volume on the headset is on all the way, the student places a hand in front of his/her mouth in front of the mouth piece and blows at his/her hand. If the student hears the wind hitting the mike, the mike is adjusted so that the wind misses the mike. Then he/she does it again.

Some "experienced" controllers are resistant to this, saying, "I've been controlling traffic for 20 odd years, and no blankety-blank kid is going to tell me the proper use of a microphone."

TYPICAL ERRORS AND THEIR SOLUTIONS

Mis-recognition of the Call Sign, Correct Recognition of the Command

This is a common error with people who have not used a radio very often. The student is usually keying the mike late. Using the foot Push To Talk (PTT) switch can help reduce this error, or telling the student to wait a heart beat after keying the mike before speaking will generally solve the problem.

The Last Word Spoken is not Recognized

This is referred to as "clipping". It occurs when PTT usage is the culprit and generally, the student unkeys before finishing the transmission. Have the student keep the PTT down for a heartbeat after finishing the transmission.

Good Recognition becomes Bad Recognition after 2 Weeks in the Cab

This is usually the result of training templates at a slow reading pace. The student may need to train new templates at a faster rate of speech.

Bad Recognition for the First Several Commands while in the Cab.

This can be a noise calibration problem. Noise calibration problems come in 2 types: 1) The student's breath hits the mike during the noise calibration. To eliminate the problem, emphasize to the student the importance of silence and stillness during calibration; and 2) The student's breath is hitting the mike during transmissions. Simply reposition the student's mike so that "P's" and "T's" do not send puffs of air onto the mike. Did you do a *MIKE CHECK*?

Inconsistent Recognition Results

The student gets a lot of commands recognized correctly, but misses a large percentage of commands. This is usually caused by a student being comfortable with some *phraseology*, but not others. Typically what is happening is the student is stuttering or pausing and re-starting in mid transmission, getting a "lazy mouth," pausing with sound, or is using incorrect phraseology.

This can also be caused by a student who generally speaks fast but slows way down when he is not sure what to say or *panics* under stress. The ITT recognizer can understand speech spoken at half the speed, or twice the speed at which it was trained.

COMMUNICATION RESEARCH BENEFITS TO THE ACCIDENT INVESTIGATOR

Stephen Véronneau, MD
Federal Aviation Administration
Civil Aeromedical Institute

ACCIDENT INVESTIGATION

When walking at an accident scene, it is quite common for investigators to climb to as many vantage points as possible to survey the site. Often, it can be difficult for them to conceptualize the totality of what had transpired in those few moments immediately preceding and during the crash. The hallmark of an excellent accident investigation is the pursuit of truth and the ability to seek information from novel sources to assist in the fact finding and determination of the sequence of events. This retrospective gathering of information is painstakingly checked and rechecked for accuracy and consistency. Slowly an understanding is built of mishaps. This reconstruction has often been likened to a detective story (Barley, 1970).

Objective accident investigation requires meticulous attention to detail and the avoidance of premature analysis particularly in fatal accidents where crew members' testimony cannot be obtained. Team members must always remember that the perishable nature of the evidence and the technical challenges of the reconstruction demand a dogged persistence of fact gathering even in situations which at first seem very distant to the cause of the mishap.

Inconsistencies and problems with reconciliation of disparate facts is not an unwelcome occurrence in accident investigation and can serve as a breakthrough in the case. In any given accident the source of pertinent facts can be quite unexpected. Occasionally the source will be met with skepticism by persons who believe that there are traditions to be followed and that form must precede function. Any source of information should be pursued to at least capture information which might only later be understood in its relation to causation. Generally, accident research has been hindered by inadequate access to the facts, circumstances, and unsummarized details of the mishaps. In the aviation environment access has been exemplary, due perhaps to the keen scrutiny all aircraft accidents receive.

THE DIFFICULTIES

One of the interesting operational limitations of communications in the aviation environment is that pilots, air traffic controllers, and dispatchers cannot use the non-verbal body language cues as extra linguistic sources of information when communicating with one another. Some say that these cues may comprise some 60% of the information content in human communications. The inability to see each other has been partly offset by the use of standard procedures and phraseology; but, it is very common for misunderstandings to occur. Recent studies of altitude deviations reported in the NASA Aviation Safety Reporting System (ASRS) report that 80% of such deviations arise from communications difficulties.

To the investigator of aviation accidents and incidents, the potential source of information from the various communication modes cannot be overlooked. Persons with operational experience evaluate the recorded communications as a measure of pilot performance against the standards and procedures of training and policy. Psychologists evaluate the emotive content and physicians may be interested in the communications as a record of possible impairment or incapacitation. Still, it is often necessary to enlist the interpretation of persons familiar with the crew, such as co-workers, close friends, or family members to assist the investigators in interpreting the nuances of the recorded communications post accident. When surviving crewmembers are available, it is interesting that they are able to provide comments and corrections to the official interpretation of the cockpit voice recorder. To this end there is a need to develop metrics of the various communication modes to assist in quantifying and understanding components of voice communication which, although audible, remain as subjective impressions.

To assist in the measurement of voice communications, the effects of the aviation environment upon voice quality need to be studied. The effects of medications, ambient temperature, hypoxia, decreased air density, fatigue, vibration (particularly rotorcraft) in addition to the stress arising from the situation or emergency would be of great help to mishap investigators. Voice stress analysis, as employed by the NTSB, has proved useful in accident investigations in several transportation modes. Voice stress analysis, in conjunction with information gathered by traditional accident methods, was very helpful in a general aviation accident where a heart attack in-flight precipitated the crash.

SOME REMEDIES

Recent accidents have demonstrated that the complexity of modern aircraft can lead to very difficult or unsolved investigations. This is particularly true when certain key data cannot be reconstructed or differing accident scenarios cannot be resolved. There have been proposals which suggest that the capability of cockpit voice recorders may be enhanced by improving the fidelity and length of the sound recording. Often when the area microphone is the main source of input for the recording, there are many more ambiguities than when boom microphones are used. Boom microphones record close to each person's mouth and all voice communications are recorded on a discrete channel. These cases would benefit from increased gathering of information by expanded digital data flight recorders.

It also has been proposed to place a video camera with a wide view of the cockpit to gather essential non-verbal information in the event of a mishap. Quick access recorders and flight data recorders are routinely accessed after each flight to evaluate operations and improve safety, without penalizing aircrew members, by some airlines in other countries. New methods of handling air traffic control radar data have been developed to provide greater insight into the view of the incident, accident or deviation from the ATC perspective. Synchronous replay of the ATC radar display data with air to ground voice recordings would also be beneficial to the investigation team.

In addition to the move to acquire greater amounts and types of flight information, we also should push for enhanced methods of analyzing and discriminating the content of the various aviation communications modes to improve mishap investigation and flight safety. At the symposium on the Methods and Metrics of Voice Communication I presented the recent history of chaotic systems research. Systems in which small differences in initial state lead to vastly different outcomes, without displaying damping of the small initial differences, is a characteristic of chaotic systems. Chaotic systems research has demonstrated some utility in ship capsize accidents by examining dynamic stability versus static measures of stability, heart rate variability and predictability post myocardial infarction, and cardiac and brain wave pattern analysis. Recent research into complex dynamic systems has produced several innovative approaches to analyzing systems with non-linear components.

Voice communication is a highly non-linear system which might benefit from an application of a non-linear systems theory. For example, chaos theory might be applied to study previously unassailable problems and wavelet applications might replace traditional Fourier transformation in speech research (Kadambe & Boudreaux-Bartels, 1992). A possible initial application of chaos theory in aircraft accident research would be to study passenger-passenger time differences in exiting aircraft. The traditional flow rates through various exits, with and without decreased visibility and with varying seat pitches do not adequately describe the flow characteristics of a group of discrete individuals moving as a type of fluid out of various sized and accessible apertures in the aircraft to the outside.

REFERENCES

Barley, Stephen. (1970) *Air Crash Detective.* New York NY: Morrow.

Kadambe, S. and Boudreaux-Bartels, G. G. (1992). Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans. on Inf. Th.,* Vol. 38, No.2, pp.917-924.

SUBJECTS CONCERNING AUDIO TAPE ANALYSIS

Larry Porter
Lawrence L. Porter Associates

INTRODUCTION

I'm an ex-air traffic controller, retired, and the answer to the question I invariably get asked, is No! I didn't get fired in the strike. I resigned in 1979. Strictly a personal decision. As a gentleman said here this morning, I spent 4 years in the Air Defense Command trying to run aircraft together, then 22 years in the FAA trying to keep aircraft apart. So much good groundwork has been laid here already this morning, I'll just jump right in here and try to keep this almost as short as Martin did. I'm going to talk about 2 things primarily: 1) determining the methodology and 2) situational awareness. Sometimes these 2 things overlap for me. Determining the methodologies to be used on a particular tape is something that I have been doing for the last 13 years, which involves enhancing audio tapes by filtering out noise and trying to enhance the speech. The second area I am going to talk about is situational awareness which includes such areas as the Air Traffic Control System, the military control room environment, and 911 calls received at police communication centers. Basically, what it boils down to is situational awareness from the standpoint of: Where are all these voices coming from? What's the network setup here? How can you keep from getting the voices from all these sources intermixed and confused when you're doing your work? I have 2 examples that I feel are interesting, from my work in this area.

Determining the Methodology

The first one I'm going to show you doesn't have anything to do 'per se' with speech, although it was a voice tape. It was a 1/2-inch multi-track tape that was alleged to have been tampered with, and I was asked to take a look at it. From Exhibit 1 you see what was thought to be a spliced out section that was only 2 1/4 seconds, but actually as you'll see later, it was 13 seconds. I physically examined the tape in person and found a 1/2-inch reel to reel tape on a 10 1/2-reel that had obviously been broken and spliced. It was a very crude splice and had about an inch and a half of clear tape wrapped around it, with the 2 ends butted up, and each end was folded over a little bit. It was pretty obvious that the tape was spliced. After looking at the tape, I noticed some interesting things about it, so I made a re-recording of the multi-track tape and, while I was at it, the digital time code channel on the tape. If you are not familiar with the sound of digital time code, it's kind of a low rhythmic thumping sound on one of the channels of the multi-track. It actually has its own rhythm that you can hear. So I recorded the voice content of the speech on the left channel, and on the right channel I recorded the digital time code. I took it to my friend, Dr. Alan Reich, and he ran a spectrogram on it for me (Exhibit 1).

The area in question is the area of the splice that you can very lightly see where the leading edge of the tape passed the head, and where the trailing edge passed the head. In the lower left, you can see where a word ends abruptly. In the same area, after the splice, we have the end of another word that doesn't tie in, or make any sense. So it looked as though we were dealing with a 2 1/4-second segment of what was obviously a gap in the tape. To verify this, I went over to the engineering lab and on a computer there created the picture in Exhibit 2. This picture is just a simple wave form which verifies that there was a gap in the tape. Then I used another program that would help me determine almost exactly what the length of that gap was which is displayed in Exhibit 3.

You can see the 22,376 points difference between the 2 cursor points which at 10,000 points per second give 1 about 2.24 seconds. Then, since I could hear the time codes so distinctly, I thought why can't you get a picture of these things? So, I wrote the manufacturer, and got a printout of the format of how the time code was constructed, which included how the pulse groups were constructed. With this information, I was able to amplify the time code data into a full wave rectification to increase its strength and then put it through a low pass filter (bandpass of 30 Hz) so that the format would resemble the one that the manufacturer had furnished. The resulting pictures are Exhibits 4 and 4a.

Each one of these spikes is 1/10th of a second in duration. Using the "P Zero" and "P Reference" points, and the fact that the standing wave is at least 2 to possibly 3 times greater than the width of the 1/10 of a second spikes, one can measure the elapsed time. Each standing wave that is wider here has a numerical value which allows one to determine hours, minutes, and seconds. In this case it's really only minutes and seconds. Exhibit 4a shows one that represents hours. In Exhibit 4, the waves that are assigned 10 seconds and 20 seconds are what I just called standing waves, or standing spikes. By adding 10 and 20 we get 30 seconds. The same thing applies to the next pulse group representing minutes. In this case, you can see it's 2 and 10, or 12 minutes. The same thing applies to the hours. So the incident occurred, or rather the drop-off of the time code occurred at approximately 9 hours, 12 minutes, and 40 seconds as one can see in Exhibit 5. What I had to do was look ahead of that time, and

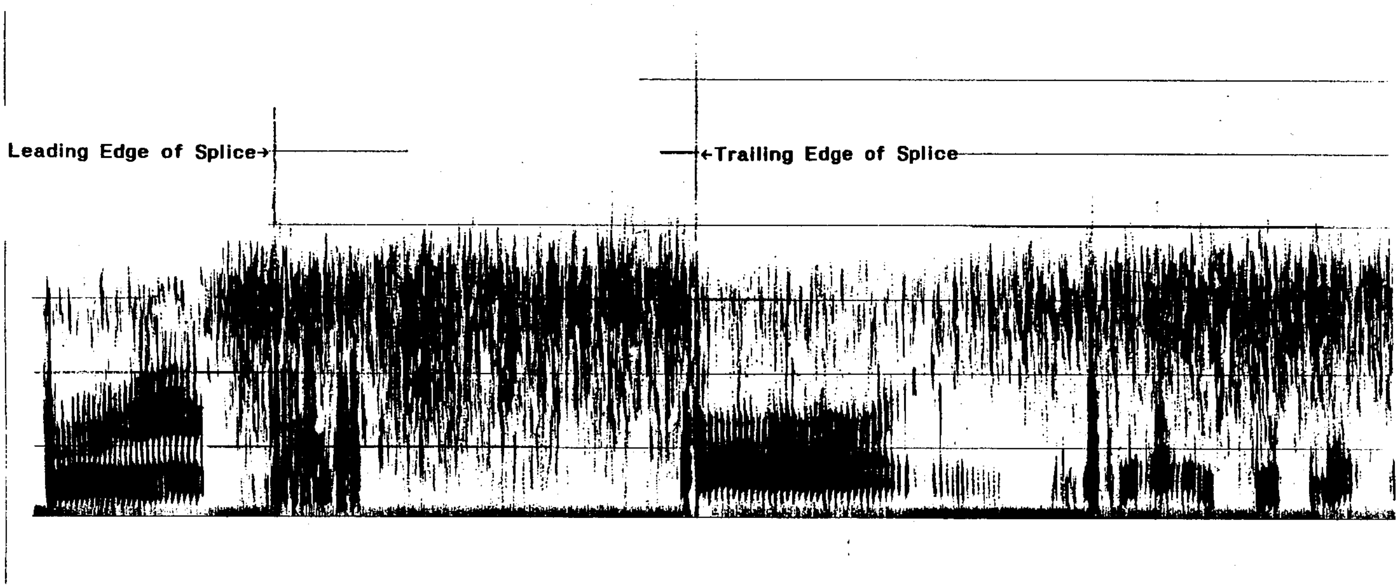


EXHIBIT 1

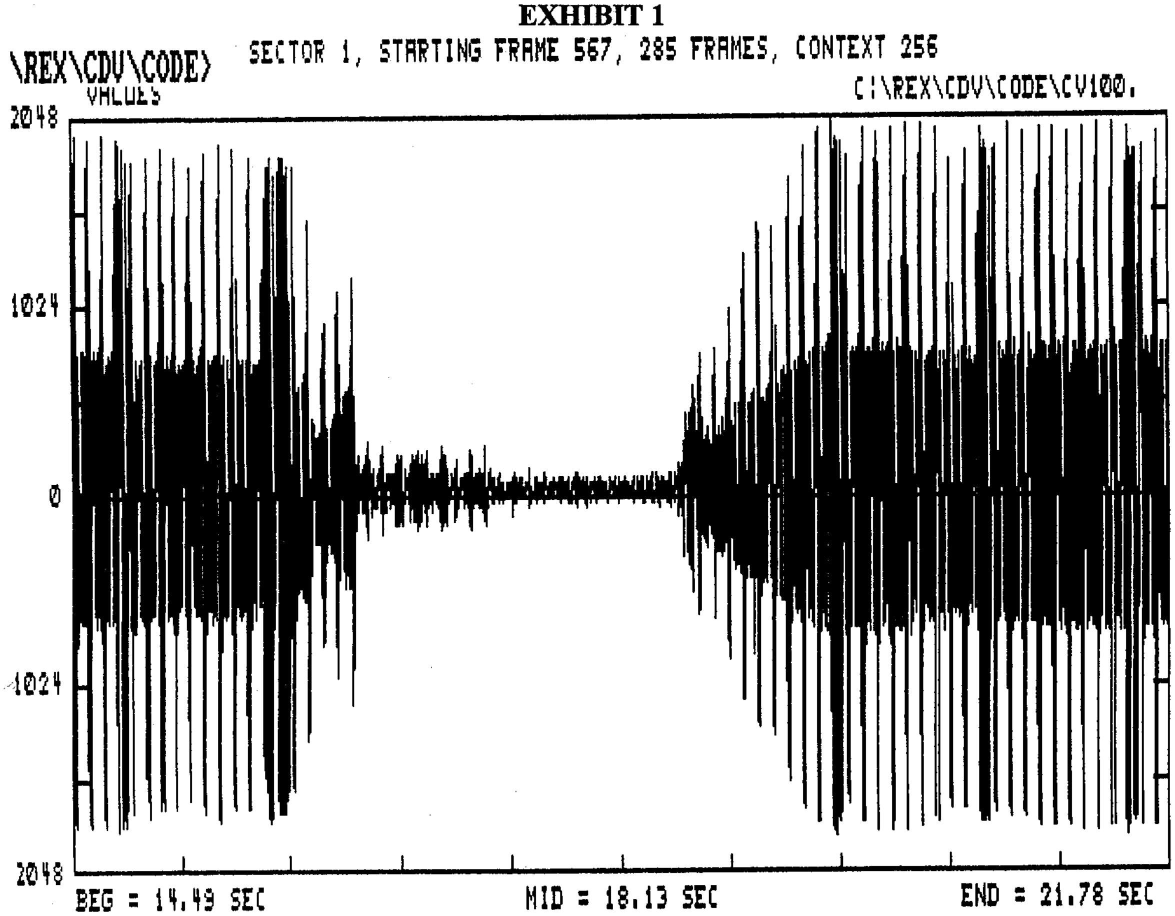


EXHIBIT 2

SAMPLE RATE 10KHz
10000 PTS./SEC.
1 PT. = .0001 Sec.

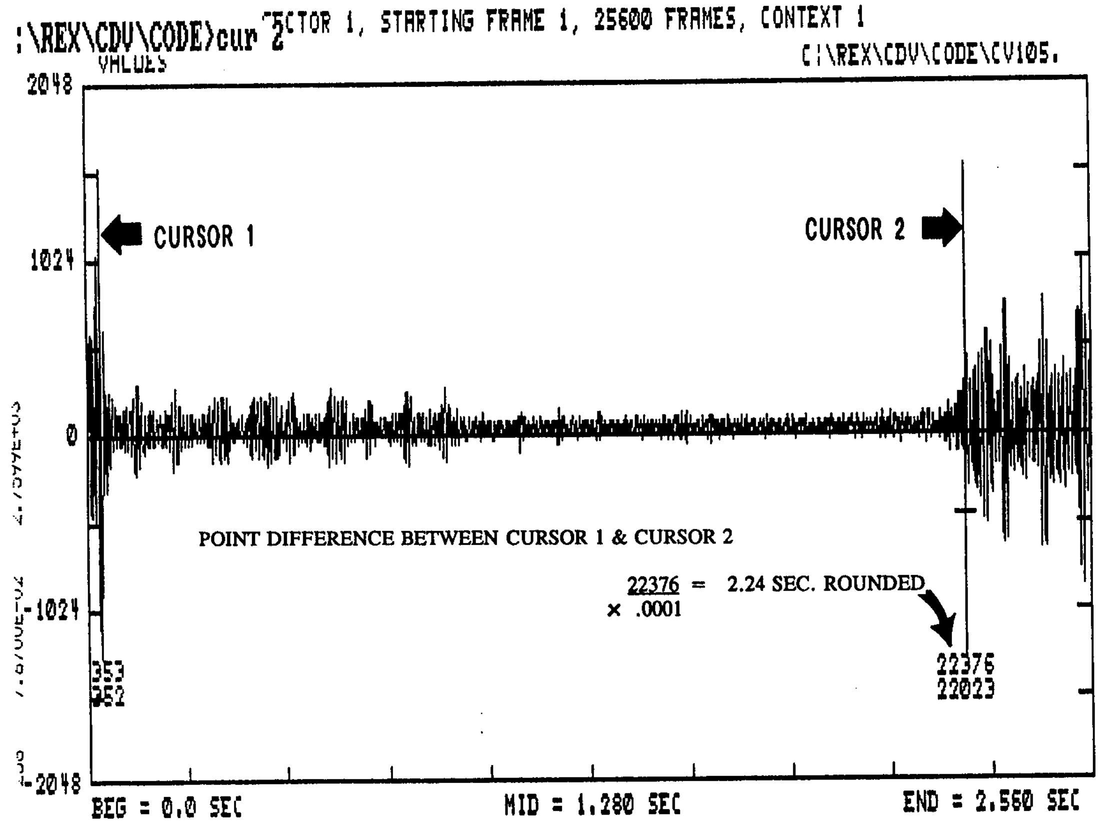


EXHIBIT 3

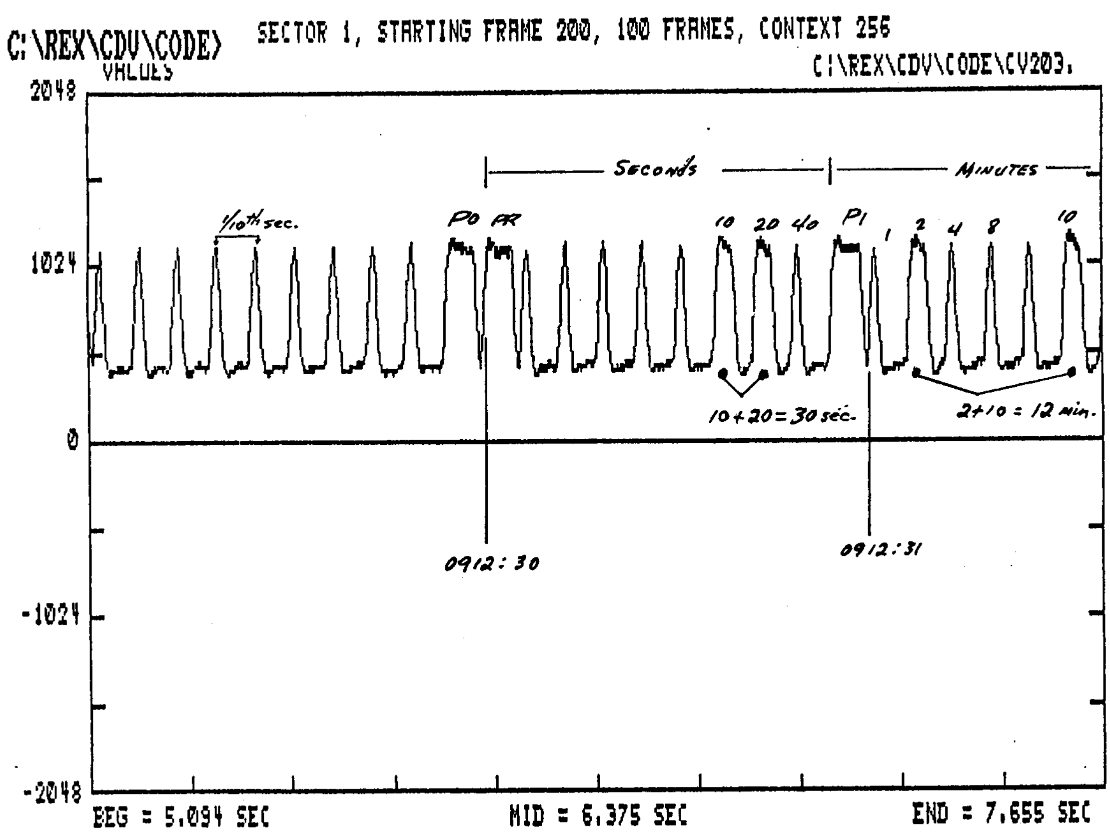


EXHIBIT 4

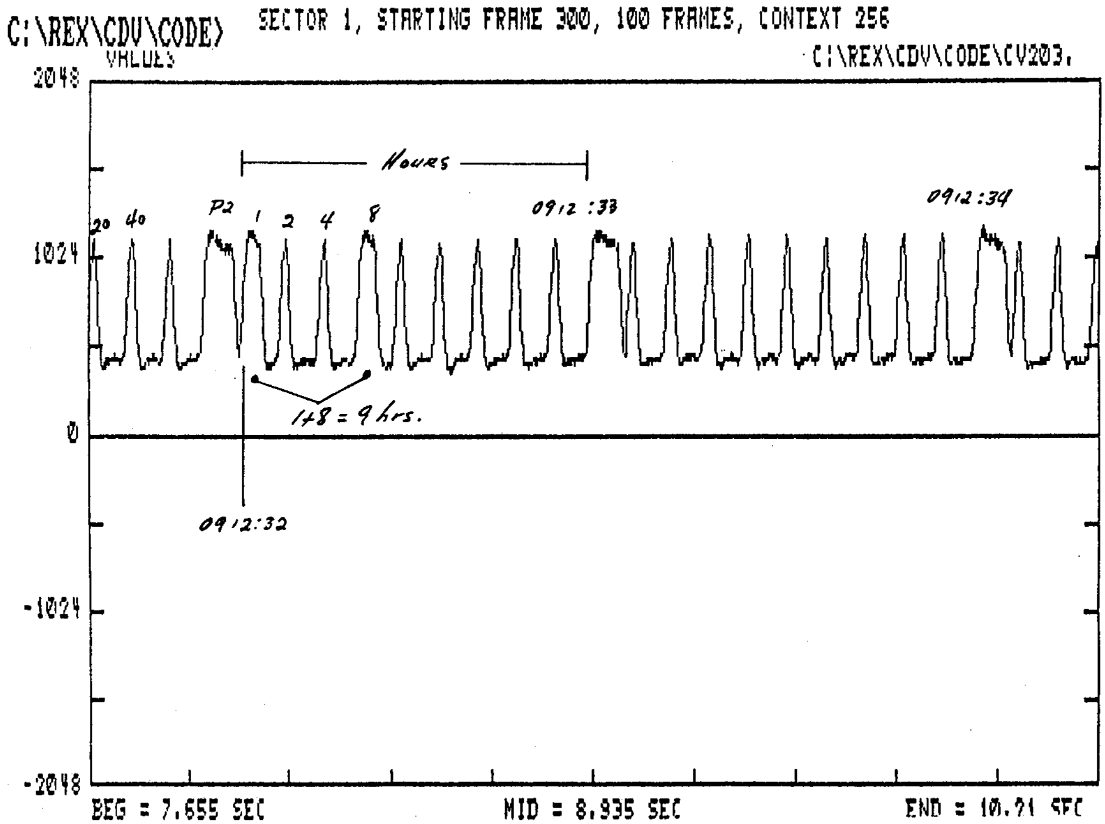


EXHIBIT 4A

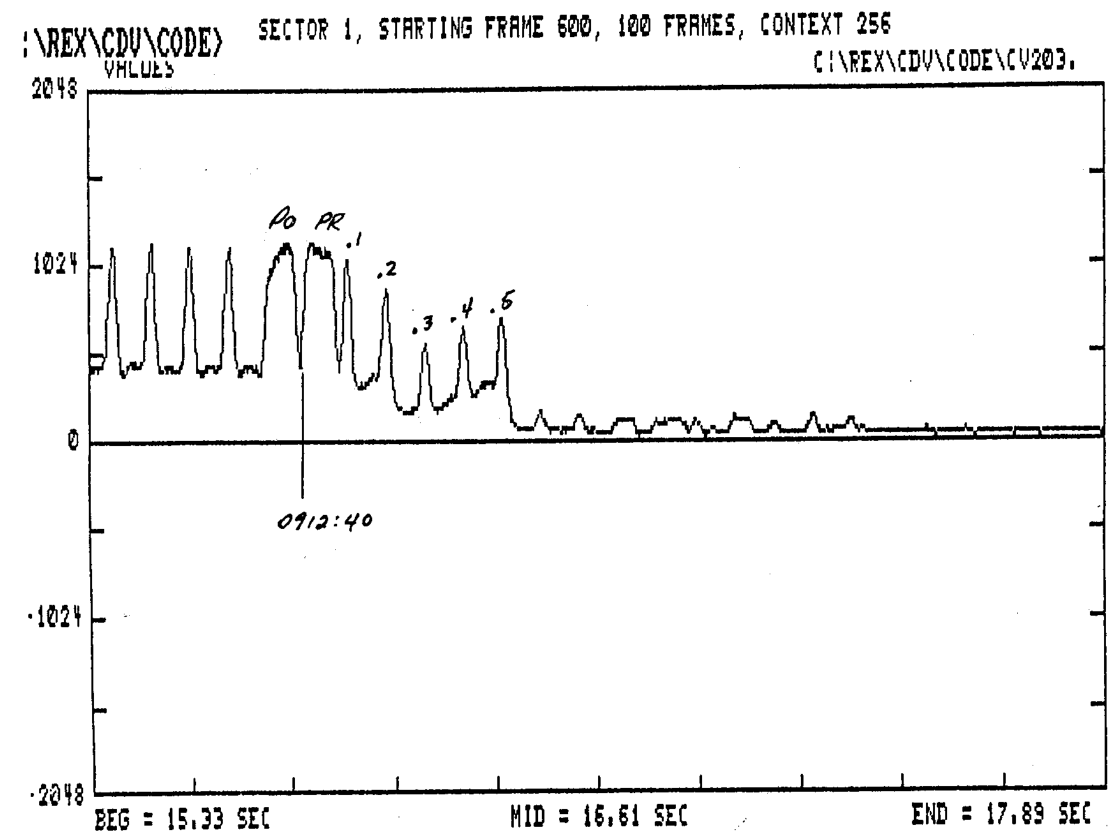


EXHIBIT 5

then look behind it also. That is how I figured out what was going on. All you'll see are 1/10 of a second spikes until we get to the break point. Here we come to a 9-hour, 12-minute, and 40-second period.

This is approximately where the tape splice was, and you can see that the time code starts to decay there, and so this is where it dropped off. From the wave forms, and things before, we knew we had a 2 1/4-second gap. Exhibit 6 shows where the time code came back to full strength at 9 hours, 12 minutes, 56 seconds. So using the same method, we got to another full pulse group at 9 hours, 13 minutes, 00 seconds (Exhibits 7 and 7a).

What was interesting about this was that it actually showed that when the time code resumed there was actually 12.56 seconds of tape missing. The multi-track tape travels at .47 inches per second, so that represents just under 6 inches of tape that was missing. So, since we had the obvious splice, that was one thing. I could not believe that at .47 inches per second you could break that tape by playing it back and forth, because it has a brake on the drum, so that if the tape does break, the bar drops down, and it keeps the end of the tape on the drive drum from slapping around. What this shows is that the tape wasn't broken just once, but twice. In my opinion, 6 inches of audio tape were missing for whatever reason and the persons involved admitted they broke the tape. I just don't buy breaking the tape twice. This is just one example of a type of methodology that I employ to visually display the precise time at which an event occurred which usually included tapes with audio.

Situational Awareness

The second part of the talk addresses the issue of situational awareness. For example, if an incident occurs in the Air Traffic Control System, it often involves more than 1 working sector or control position and sometimes more than one facility. If you make a request for information on an accident, you usually get only a tape of the last person that had contact with the aircraft, and sometimes that's not enough to go on. This first example involves an incident at a major airport with a pilot and several radar personnel: 2 radar controllers and a data controller. I'll give you just a short background. A light aircraft departed a satellite airport, headed westbound, and got about 28 miles west of the major airport (where the air traffic services and facilities are located) where he encountered some fog. Basically, he got himself into instrument conditions. The approach controller who was providing radar vectors asked the pilot, "Do you have visual contact with the ground?" In listening to the tape of that radar working position, it sounded like the pilot said "affirmative," but there was just enough of a problem right in this area that it caused me to wonder. In addition to the radar controller, there is a data controller who handles most of the coordination and paper work. This person is there basically to assist the radar controller. In addition, both the data position and radar position have a set of hotlines and the data position also has a set of interphone lines right in front of him. The interphone position has a flip-flop toggle override switch so that he can plug into a jack on the other side of the room and still be able to monitor the same hotlines and radio channels as the radar controller. There was also a third radar position just to the right of the radar controller's radar scope, but it was not staffed at the time so the tape of that station provided another tape of this communication. I made a re-recording of the third radar position tape and data position tape and compared those with the tape that I'd been furnished. What it turned out to be was that a tower controller at another airport initiated a call on the hotline ("Approach, Tower...") right after the radar controller had asked this question of the pilot: "Do you have visual contact with the ground"? Since the tower had initiated the call at the precise time that the pilot started his response, the radar controller's position recorded an "Ah" sound right there (indicating the "A" sound) which was presumably the pilot. At that same time the interphone controller punched the hotline to intercept the call essentially cutting that word off (which is represented by "####" in Exhibit 8) which created a disturbance over that part of the pilot's response. The interphone hotline disconnects the radar controller so that he doesn't have that coming into his ear.

Playing back the third radar position, which has the same radio frequencies recorded on its channel, one can tell what was said. The radar controller said, "Do you have visual contact with the ground?" and the pilot definitely said, "Negative." That makes a very big difference. The pilot went on to state that he was flying straight and level, heading 280 degrees at 2,900 feet and a speed of 120. The only variable in this transmission that I don't think we'll ever know is what the controller heard because the controller subsequently took no action to help the man. I guess that will never be known.

The second example of situational awareness issues involves a cockpit voice recording tape of a DC9 that crashed in Detroit due to windsheer. What was interesting about this, which is something that I encounter frequently, is that sounds or voices from more than one source will suddenly intermix together to form something that you know you heard. I was asked by another consultant to take a look at this tape and attempt to run some techniques on it. However, the tape that I received was recorded in mono. Originally, it would have been possible to record the cockpit area mike (CAM) on the left channel, and the ATC communications on the right channel. However, in this case, somewhere in the chain of recording, and

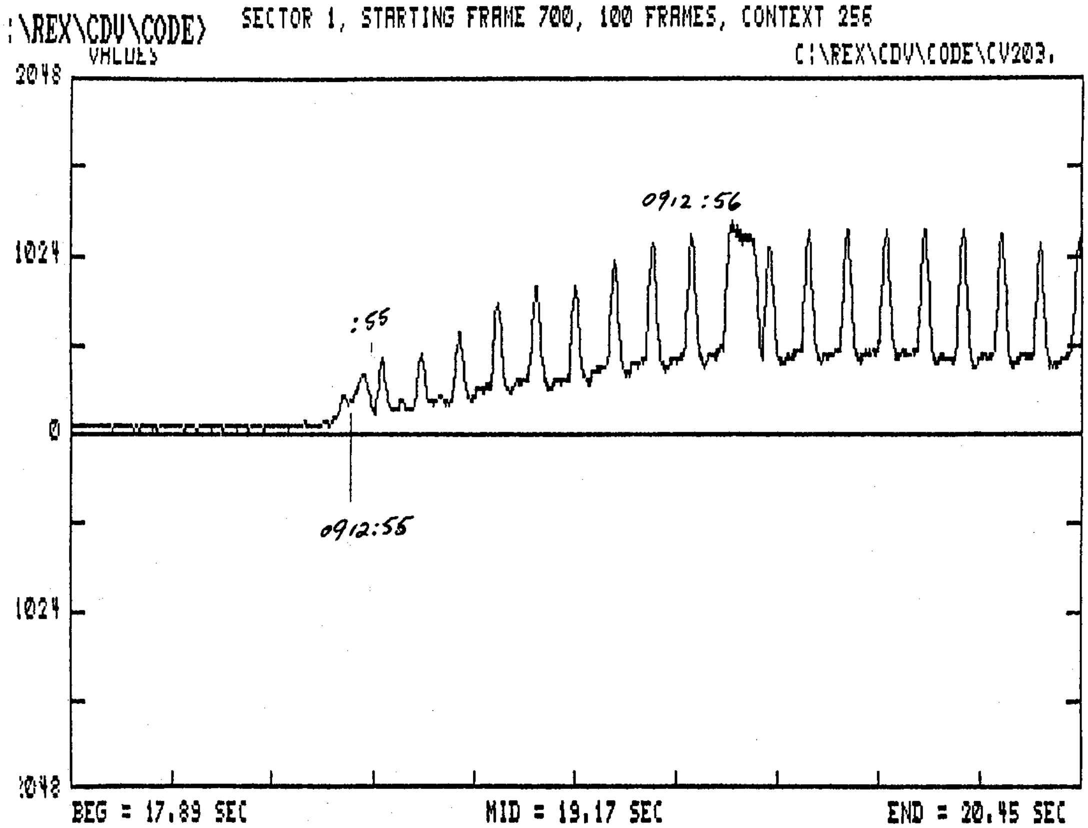


EXHIBIT 6

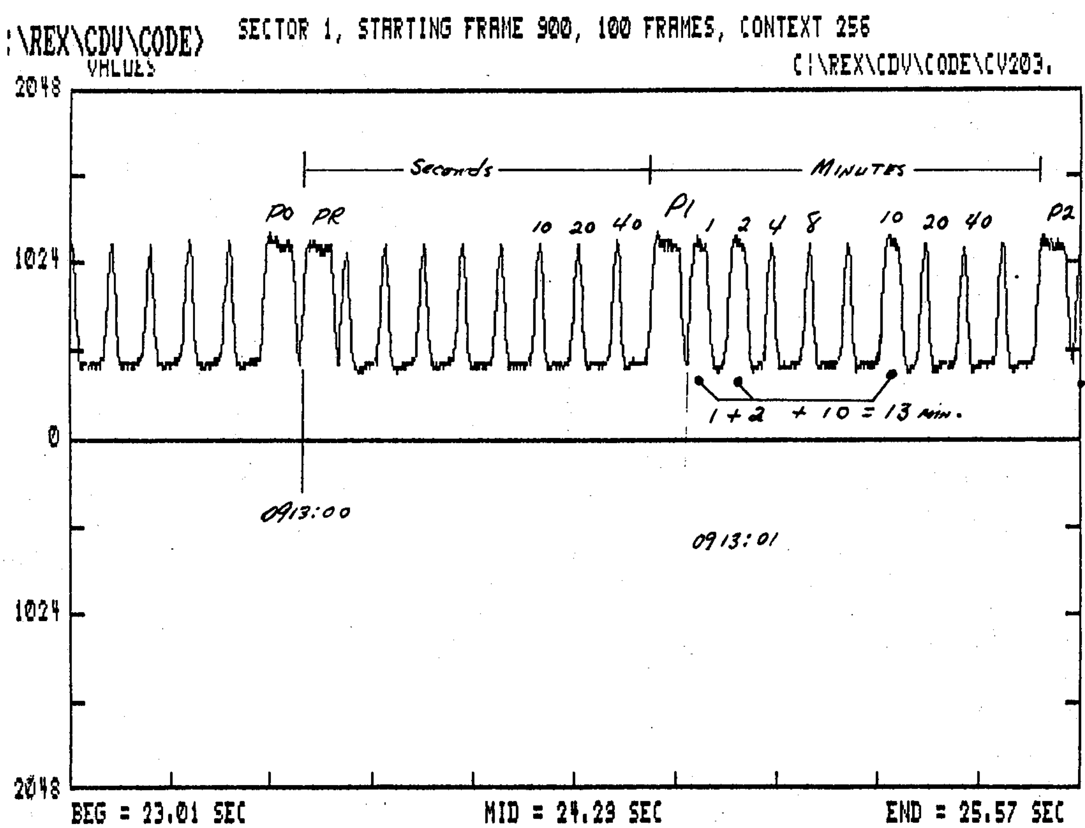


EXHIBIT 7

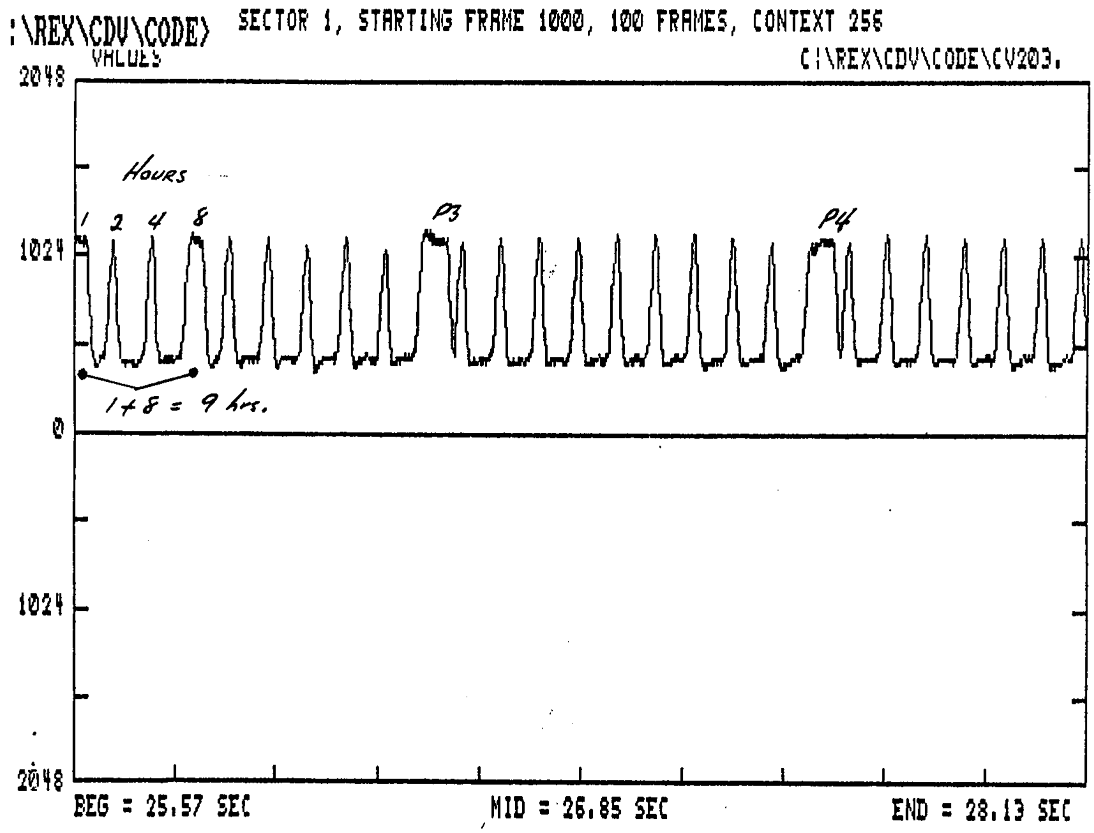


EXHIBIT 7A

SITUATION 1:

CNTLR: DO YOU HAVE VISUAL CONTACT WITH THE GROUND?

AIRCRAFT: A #### TIVE

(Tower) (noise) (end of response from aircraft)

The comparison of another channel of the same tape which was free of interference showed that the pilot's response was actually the word "Negative".

SITUATION 2:

PILOT: WELL I'LL BE DAMNED

CONTROLLER: FRONTIER.....

The words "well I'll be..." were much softer than the emphasis on the word 'damned'. The beginning of the controller's transmission, "Frontier..." combined with the pilot's emphasis to form what sounded like 'Down the gear'.

EXHIBIT 8

re-recording, someone had re-recorded both channels on a mono system, and thereby essentially blended both channels together. By the time I got the tape, I couldn't separate it, so I had to work around it. The crew was running the aircraft with the speakers on instead of wearing their headphones, and the speakers were very loud. There was a lot of thunderstorm activity that night, and the Air Traffic Controllers were really up on a step. Their voices were coming in real loud on the speakers, so I could actually hear the ATC communications better than I could hear the 2 pilots. The remark in Exhibit 8; "Down the gear," is what was on the transcript of the company tape that I received. I was able to remove or de-convolute most of the distortion and sudden interference that was saturating the cockpit area microphone. What actually was said during that time was, "Well, I'll be damned." That was said right before it dropped out from under the pilot. These three words here ("Well, I'll be ...") were softer, then it kind of built up with a lot of emphasis on the word "damned," but at the same time the controller was initiating a call to a Frontier 214. So you have the word "damned" and the word "Frontier," and it comes out "down the gear", because the words smashed together, but we were able to separate them a little bit. The phrase didn't make any sense to me because I'd heard the call for gear down; heard the response; heard the lever activated; and heard the gear come down. That was my problem with that phrase, because it didn't make any sense to me. I didn't go into this project looking for something different. If for some reason they had picked the gear back up, and it wasn't down again, that's fine, but objectivity's very important in this area of work. So those are the types of things in which I became involved. The majority of my work is in law enforcement tapes and 911. I also get work from people recording their husbands and wives. I had a fellow that thought he'd caught his wife cheating on him. I showed him that it was bleed through from the back side of the tape because he has a party line and that the voices he heard were actually 2 of his neighbors talking. I thought he was going to be happy, but he was absolutely furious with me.

Old visual representations of time code information, such as the linograph, present nothing but a wave form. The approach just shows you that the wave, or the time code did exist, and then it didn't exist, but that is all it tells you. It doesn't indicate to me exactly what time it is. And I think that's important. That's all I have.

SPEECH ANALYSIS IN RUSSIA

Malcolm Brenner, David Mayer, James Cash
National Transportation Safety Board

INTRODUCTION

This paper resulted from the May 1989 Agreement on Cooperation in Transportation Science and Technology between the United States and the former Soviet Union. As part of the original agreement, a subgroup for Aircraft Accident Investigation was formed. The National Transportation Safety Board (NTSB) and the GOSAVIANADZOR of the Soviet Union began cooperative technical exchanges of specialists and material related to accident investigation and prevention. Following the 1991 breakup of the former Soviet Union, the cooperative exchanges continued between the NTSB and the newly formed Interstate Aviation Committee (MAK) that represents the accident investigation authorities of the Commonwealth of Independent States (CIS). This paper resulted from a continuation of the cooperative work of the Accident Investigation Group.

There has been an exchange of papers and personal visits related to areas of scientific cooperation, exchanges of the sort that were not possible during the political climate that prevailed between our countries in most of the recent past.

In line with this effort, our agency provided information to our colleagues in the CIS concerning speech analysis work that was accomplished by our staff (Brenner, M., & Cash, J.R., 1991; Brenner, M., Doherty, E.T., & Shipp, T., 1994). In return, we received a remarkable letter from Alfred Belan, M.D., chief of the acoustics laboratory of the Interstate Aviation Committee in Moscow. The letter, written in broken English, claimed an ambitious program of speech analysis work of which we were completely unaware. The letter indicated that Dr. Belan was preparing a book in Russian describing observations made from the speech recordings of more than 300 airplane accidents. It should be noted that there are perhaps 30 airplane accident voice tapes discussed in English-language articles (Ruiz, R., Legros, C., & Guell, A., 1990). The letter, then, suggested a level of experience that was an order of magnitude greater than that of the entire scientific literature! Intrigued, we invited Belan to visit the United States for further discussions.

In February 1994, Dr. Belan spent a one-week visit at the NTSB headquarters in Washington, D.C. In addition to our staff, Barbara Kanki of NASA-Ames Research Center attended the meetings. The meetings consisted of both discussions and laboratory analysis of accident tapes.

Dr. Belan was a pleasant man in his late fifties, highly educated, who spoke little English but displayed a clever and charming sense of humor. Some of the credibility assigned to the Russian research came from the very favorable impression made by Dr. Belan himself, especially given the inherent language difficulties.

The information described in this paper is based on our meetings with Dr. Belan. This represents our best, albeit limited, understanding at the time of the Russian program.

Origin of the Russian Speech Analysis Program

The Russian speech effort began about 20 years ago and was centered in the Institute of Aviation Medicine. The work was inspired by the 1969 paper of American researchers Williams & Stevens (Williams, C.E., & Stevens, K.N., 1969). Early work from the Russian program was published in English (Simonov, P.V., & Frolov, M.V., 1973; Simonov, P.V., & Frolov, M.V., 1977). However, after the late 1970s, the work was no longer published outside Russia and it apparently took on something of a secret quality. Speech analysis was used to evaluate cosmonauts and pilots for fitness for duty in terms of both stress, fatigue, and other aeromedical qualities.

The program used simulator research, in some cases with test pilots as subjects, and also studied pilots and cosmonauts during real life aerospace situations. In the case of fatigue, for example, subjects performed in research projects for periods of 72 hours without sleep. Fatigue studies were made of cosmonauts in extended duty situations. In addition to research, systematic examination was made of aviation accident tapes from both military and civilian accidents.

Measures Used in the Russian Research

Dr. Belan referred to numerous speech measures used in Russian research. Although some were new to us, many were familiar from English language literature. What was striking about the Russian approach was its broadness, combining acoustic, phonetic, and communication information in a way that seemed original. What was also striking was the seriousness with which the measures were applied and the level of experience shown with the measures.

The Russian effort groups speech measures into 4 categories, which are evaluated for each speech sample. The categories are:

1) *acoustic measures*. These include fundamental frequency; fundamental frequency range; amplitude; and relative energy distributions among the formants. The last measure was of special interest, following from early work published by Russian researchers (Simonov & Frolov, 1973, 1977). At least some of these measures are extracted by automated techniques.
2) *timing measures*. These include speaking rate, and measures such as relative speaking/silence time and latency to respond.
3) *contour measures*. These relate to the relative shape of the speech energy waveform when plotted over time. An example would be whether the waveform is relatively flat or spiked.
4) *psycholinguistic measures*. These include phonetic measures such as changes in articulation of words. They also include measures of communication, such as whether communication is appropriate and effective given the ongoing conversation and the demands of the flight situation. One of the most interesting aspects of the Russian work is that it formally compares evidence based on the physical properties of speech with evidence based on the effectiveness of communication.

Proposed Standards

Based on his experience, Belan suggested general standards that apply to normal human speech. We have not seen such standards published and found them immediately practical in our work. We report them here for review by our colleagues.

For fundamental frequency, Belan suggested that a male speaker engaged in relaxed communication should display an average fundamental frequency between 80-130 Hz. The range should be higher, 95-145 Hz, in cockpit situations (perhaps because the speaker is compensating for background noise). Thus, if a pilot displays an average fundamental frequency that is higher than 145 Hz, regardless of the flight situation, it is abnormal and a sign that the pilot is very tense. (Belan noted, however, that intra-individual changes are more important than absolute changes on all speech measures).

For fundamental frequency range, Belan suggested that an average range of 45-75 Hz was normal in a relaxed situation. A range of 45-90 Hz was normal for a dynamic flight situation.

For speaking rate, Belan suggested an average rate of about 4.5 to 7.5 syllables per second as normal. A phrase might contain as few as 4-7 syllables, and in some cases as few as 2-3 syllables if the words were conversational, and still provide useful data for measuring speaking rate.

For segmenting statements, Belan suggested that a silent period of 300 msec be used to delineate the end of one statement and the beginning of another. This might represent an approximate minimum time necessary for a human speaker to shift thoughts.

An Example of Russian Work: Psychological Stress

As an example of Russian work, Belan described in detail some work on the speech effects of psychological stress. He provided a lecture on this topic, and demonstrated his thinking in a laboratory analysis of several accident tapes.

In general, the Russian work discusses 3 stages in the human response to psychological stress. These range on a continuum from a constructive response to absolute panic. The stages can be characterized as follows:

Stage 1. Belan described the first stage of stress as a working stress that improves performance, a constructive mobilization of attention and resources in reaction to an unusual event. The speaker is in control of speech, communications are accurate and there are no logical or semantic disturbances evident in speech. The pilot's performance in the cockpit shows no procedural errors. In acoustic and rate measures, this stage is characterized by an intra-individual increase of about 30% in fundamental frequency when compared to relaxed levels, an increase of about 10% in amplitude, and, perhaps, an increase of 5-10% in speaking rate.

Stage 2 . The second stage of stress was described as just strain. The pilot can still do the job and make decisions. Movements can become sharper but are still under control. The pilot does not make gross mistakes.

In the second stage of stress, speech is still adequate to the situation but emotional stress is clearly seen. Speech is fast, strained, brief, and accented. There may be a reduced latency to respond (such as the speaker's response beginning before the query is complete). Occasionally, phrases are not completed. Belan noted that there is a reduction of nonessential speech: the speaker "observes the purpose of communication." Speech may be repetitious as if to ensure that the recipient understands.

In Stage 2, the speaker's performance often displays hasty or premature actions. Intermediate procedural steps may be skipped, such as the omission of checklist items. The speaker appears to be trying to overtake the situation.

Stage 2 speech is characterized by an increase of 50-150% in fundamental frequency when compared to relaxed levels, an increase of 15-20% in amplitude, and, perhaps, an increase in speaking rate of more than 50%. Other signs of stress include an increase in fundamental frequency range and contour changes. Measures of pulse and respiration would show increases.

Stage 3. On top of all else, during Stage 3, the pilot cannot think straight. Sometimes he cannot speak clearly, leaves out letters, and repeats the same thing. Sometimes his answer is unrelated to the question. He is apparently thinking of something else. Belan says that speech is characterized by those things that dominate the speaker's thinking regardless of the situation. Standard operating procedures are not followed. There can be an occasional, stupor-like refusal to act (although this is rare).

In Stage 3, there is often incomplete articulation, with unvoiced syllables and words swallowed or not produced. There is poor word choice and improper grammar, and no attempt to correct speech errors. Fundamental frequency increases 100-200% over relaxed levels, amplitude increases 30-50%, and there can be large oscillations in rate including increases of 50-200%. Dr. Belan noted, however, that these changes may not apply to the highest levels of Stage 3. It is not unusual to see a sudden drop in fundamental frequency and hoarseness when the speaker faces imminent death.

Other Applications

Belan indicated that Russian work has examined fatigue and hypoxia effects on speech, areas in which there is no literature in English language journals. There is also work published in Russian on the physiology of physical effort and its effects on speech. These areas were discussed only briefly in our one-week meeting, however we received an impression that Russian work in these areas was as thoughtful as the work on psychological stress.

Future Directions

The Russian work appears to add significantly to previous work published in English language sources. It adds confidence that there may be characteristics of human speech that are cross-cultural and that will allow us to identify and quantify emotional responses. The leadership of the NTSB and MAK plan to continue the support of the cooperative exchanges of technical information and specialists in the field of accident investigation, and we anticipate further exchanges with the Russian program that can lead to a more involved cooperative work.

ACKNOWLEDGMENTS

We thank Eugenia Bernstein and the Russian Embassy for translation assistance during the visit of Dr. Belan. The opinions expressed in this paper are those of the authors and do not necessarily reflect the official position of the National Transportation Safety Board.

REFERENCES

Brenner, M., & Cash, J.R. (1991). Speech analysis as an index of alcohol intoxication — the Exxon Valdez accident. *Aviation, Space, and Environmental Medicine*, 62, 893-8.

Brenner, M., Doherty, E.T., & Shipp, T. (1994). Speech measures indicating workload demand. *Aviation, Space, and Environmental Medicine*, 65, 21-6.

Ruiz, R., Legros, C., & Guell, A. (1990). Voice analysis to predict the psychological or physical state of a speaker. *Aviation, Space, and Environmental Medicine*, 61, 266-71.

Simonov, P.V. & Frolov, M.V. (1973). Utilization of human voice for estimation of man's emotional stress and state of attention. *Aerospace Medicine*, 44, 256-8.

Simonov, P.V. & Frolov, M.V. (1977). Analysis of the human voice as a method of controlling emotional state: achievements and goals. *Aviation, Space, and Environmental Medicine*, 48, 23-5.

Williams, C.E. & Stevens, K.N. (1969). On determining the emotional state of pilots during flight: an exploratory study. *Aerospace Medicine*, 40, 1369-1372.

DEVELOPMENT OF A SPEECH ANALYSIS PROTOCOL FOR ACCIDENT INVESTIGATION

David L. Mayer, Malcolm Brenner, and James R. Cash
National Transportation Safety Board

INTRODUCTION

Accident Investigation

Evidence provided by voice recordings is often integral to the investigation of aviation accidents. These voice tapes may be recordings of radio traffic, or they may come from cockpit voice recorders (CVRs), which store the final 30 minutes of flight-deck sounds. These recordings have long been used to assist investigators in determining what happened in an accident. Speech analysis, however, holds promise for gaining insight into why it happened. The authors hope that speech analysis techniques will lead to a better understanding of cognitive and emotional states that underlie the behavior of people involved in accidents. This paper describes an initial attempt to develop a protocol for such an analysis.

Speech Measures

Speech analysis holds promise as a technique for detecting changes that may be associated with fatigue, hypoxia, alcohol intoxication, drug impairment, physical exertion, workload demand, emotional stress, and fear (Belan, 1994; Brenner & Cash, 1991; Brenner, Shipp, Doherty & Morrissey, 1985; National Highway Traffic Safety Administration, 1989). The present work is primarily concerned with the detection of workload demand and emotional stress. Several researchers have reported success in using fundamental frequency (pitch) as a measure of stress (Ruiz, Legros, & Guell, 1990; Scherer, 1981; Streeter, McDonald, Apple, Krauss, & Galotti, 1983). Brenner, Doherty, and Shipp (1994) asked subjects to count aloud while performing a tracking task with different levels of workload demand. They found that fundamental frequency and vocal intensity (loudness) increased significantly with workload demands, and speaking rate also showed a marginal increase. These measures, along with a derived measure similar to one employed by Brenner et al. (1994) and a syllable count suggested by Belan (1994), were used to analyze a speech sample from a helicopter accident. It is hoped that this work will lead to a standard protocol for speech analysis associated with accident investigation.

METHODS

The Speech Sample

On January 28, 1980, a U.S. Marine Corps UH-1N helicopter was enroute to Redding, California, on a visual flight rules (VFR) flight plan. The captain contacted a civilian Flight Service Station (FSS) by radio to exchange routine flight information and to change his destination to Red Bluff, California. Within moments of concluding this exchange, the aircraft sustained a catastrophic engine-to-transmission drive shaft failure and began an uncontrolled descent. Evidence indicated that the transmission and main rotor blades departed the aircraft during its inverted descent. The captain declared a "mayday" to the FSS and gave an assessment of the situation and a position report. The helicopter crashed shortly thereafter killing all onboard. All radio transmissions between the captain and the FSS were tape recorded by equipment at the FSS. An analysis of this recording was performed in the CVR laboratory of the National Transportation Safety Board (NTSB). (Because it involved a military aircraft, the NTSB did not conduct its own investigation of this accident.)

Analysis Procedure

The tape recording was digitized for computer-assisted acoustic analysis using an HP9000 workstation running the Waves analysis package developed by Entropic Software. Using expert guidance (Belan, 1994), statements were defined as utterances bounded by pauses of at least 300 msec. Using this definition, the sample contained 9 statements made during routine flight, and 14 statements made during the emergency. The routine statements were spoken over 46 seconds, and the emergency statements were spoken over 38 seconds; 21 seconds separated the 2 statements. Three sub-statements or phrases were spoken under both routine and emergency conditions. Five primary speech measures were made for each statement and repeated phrase: mean fundamental frequency (f_0), fundamental frequency range (Wf_0), duration, and mean amplitude (loudness) were determined with computer assistance, and the second author determined the number of syllables by listening to the digitized sample. Speaking rate (syllables per second) and 2 other derived measures were computed later. Speaking rate was not computed for utterances of fewer than 4 syllables. Following Brenner, Doherty, and Shipp (1994), the first derived measure (D-1) was computed by summing the *z*-scores of the f_0, and speaking rate for each statement. After Belan (1994), the second derived measure (D-2) was computed by summing the *z*-scores of the Wf_0, speaking rate, and syllable count for each statement (syllable counts were reverse-scored because, unlike other measures, they were expected to decrease during stress). Three analyses were conducted using these measures: (1) a statement analysis that compared f_0 and Wf_0 for each statement, (2) a condition analysis that compared routine statements to emergency statements, and (3) a phrase pair analysis that compared the phrases that were repeated under both routine and emergency conditions. (Because the radio equipment from which the recording was made was governed by an automatic gain control system, the amplitude measures were unusable in these analyses and they are not discussed further.)

RESULTS

Statement analysis (Figure 1) presents the f_0 and range of Wf_0 for each statement. The square plot symbols indicate the f_0 for each of the statements. Hollow squares indicate the 9 routine statements; filled squares depict the 11 statements made under emergency conditions. Error bars plot the range of fundamental frequencies for each statement.

It is clear from Figure 1 that the captain's fundamental speaking frequency was elevated during the emergency compared to his speech under routine conditions. Further the growth of range under emergency conditions is striking.

Condition Analysis

During routine flight, the captain's fundamental frequency averaged 123.9 Hz. This increased to an average of 200.1 Hz during emergency conditions. His Wf_0 changed from 124.2 Hz during routine flight to 297.3 during the emergency. Both of these elevations were significant using 2-tailed *t*-tests, which were used to avoid bias despite predicted difference directions. The captain averaged 11.7 syllables per statement during routine flight, but this dropped to an average of 6.7 syllables per statement during the emergency. (Six of the captain's emergency statements contained only the 2 syllable word "mayday.") If these statements are excluded, the average for the 4 remaining emergency statements is 7.8 syllables per statement. Both derived measures increased under emergency conditions, but only D-2, the Russian-influenced measure, changed significantly. Two-tailed t-tests were performed on all of these observed differences, and the results are summarized in Table 1.

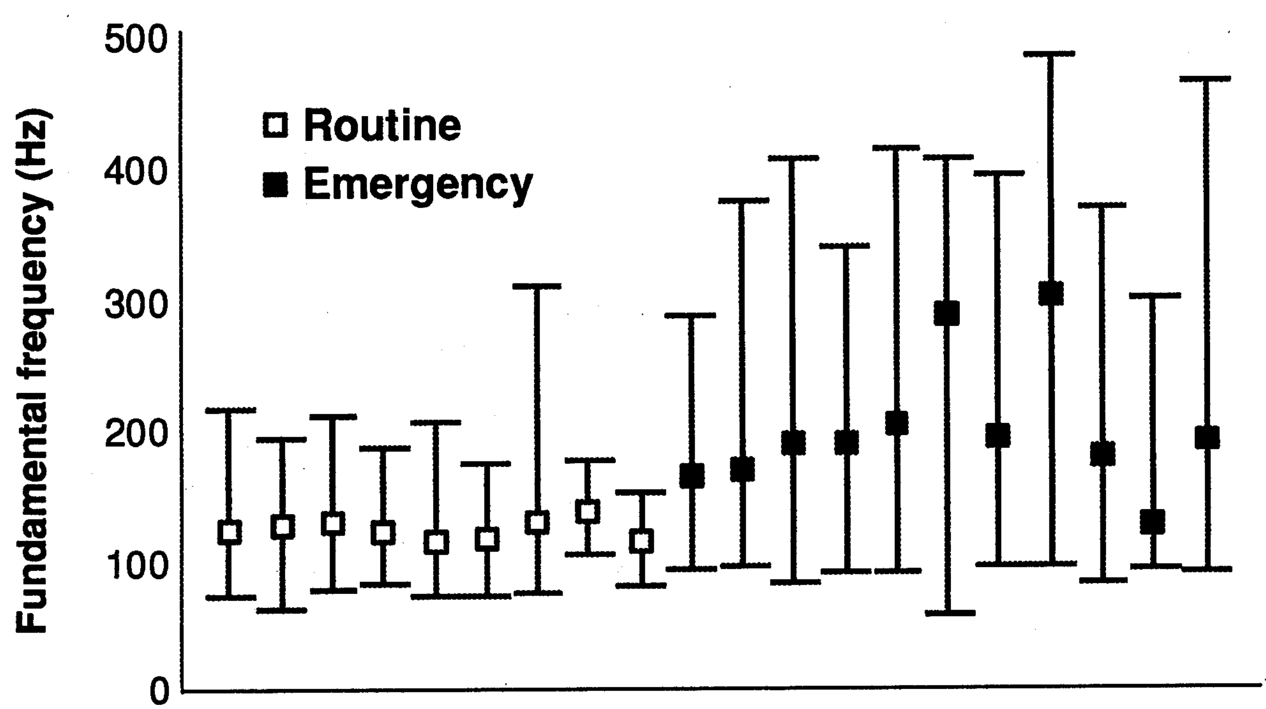


FIGURE 1: Fundamental Frequency Means and Ranges for All Routine and Emergency Statements

Measure	*Routine*	*Emergency*	*Significance*
Fundamental frequency (Hz)	**123.9**	**200.1**	**<.001**
Range of fundamental frequencies (Hz)	**124.2**	**297.3**	**<.0001**
No. of syllables	**11.7**	**6.7**	**=.054**
Speaking rate (syllables per second)	**5.3**	**4.4**	**n.s.**
Derived measure D-1	**-0.14**	**0.18**	**n.s.**
Derived measure D-2	**-0.70**	**0.90**	**<.0001**

TABLE 1: Summary of Mean Speech Measures by Condition

The information in Table 1 shows that, as predicted, both f_0 and Wf_0 increased significantly during the emergency. Also as predicted, the number of syllables per statement decreased, but this difference was not statistically significant. The derived measure used in previous work (D-1) did not change significantly, but D-2 changed dramatically. In Figure 1, the *z*-scores of the observed differences have been graphed for easy comparison. Graphical presentation of captain's speech before and during the emergency.

Phrase Pair Analysis

During the uncontrolled descent, the captain repeated 3 phrases that he had used moments earlier during routine flight. He reestablished communication by calling the FSS by its identifier, identified himself with his callsign, and gave his position. Table 2 presents speech measures for each of these phrase pairs. Although little change occurred in phrase speaking rate, large changes were seen in fundamental frequency.

Figure 2 shows the differences between the fundamental frequencies of each of phrase pairs. Each bar in Figure 2 shows the value of the fundamental frequency of one phrase, with one exception: The pilot gave his callsign twice during routine conditions; therefore, the bar that indicates this phrase actually plots the mean fundamental frequency of both phrases. A line that indicates the average fundamental frequency of all statements made during routine flight is labeled R, and a corresponding line that shows the average for all statements during the emergency is marked E (these lines plot the averages given in Table 1 for fundamental frequencies). For each phrase, the pilot's speaking pitch was higher during emergency conditions.

CONCLUSION

The extreme emotional stress experienced by the speaker during the uncontrolled descent of his aircraft is apparent in an affective sense to anyone who listens to the recording. This sample was chosen for this preliminary work because it captured 2 dramatically different emotional states, and because of the special analysis opportunities afforded by the repeated phrase pairs. The short period of time between the routine and emergency statements, and the fact that the entire recording was made using the same equipment, further made the sample attractive for this work. For these reasons, it presented a best-case scenario for development of an analysis protocol. Simply put, if the techniques described in this paper failed to work here, they would surely not work for subtler cases. The elevation in fundamental speaking frequency observed during emergency conditions is consistent with the presence of emotional stress and an increased workload demand as documented in previous studies. Further, Belan (1994) estimates that 90% of the population exhibits such a change during periods of stress.

Measure	*Routine*	*Emergency*
FSS identifier		
Fundamental freq. (Hz)	**127.3**	**193.4**
Speaking rate (syllables/sec)	**5.37**	**5.13**
Callsign		
Fundamental freq. (Hz)	**136.1**	**159.1**
Speaking rate (syllables/sec)	**6.07**	**5.38**
Position report		
Fundamental freq. (Hz)	**121.3**	**222.3**
Speaking rate (syllables/sec)	**3.21**	**4.39**

TABLE 2: Summary of Mean Fundamental Frequencies and Speaking rates for Phrase Pairs Spoken during Routine and Emergency Conditions

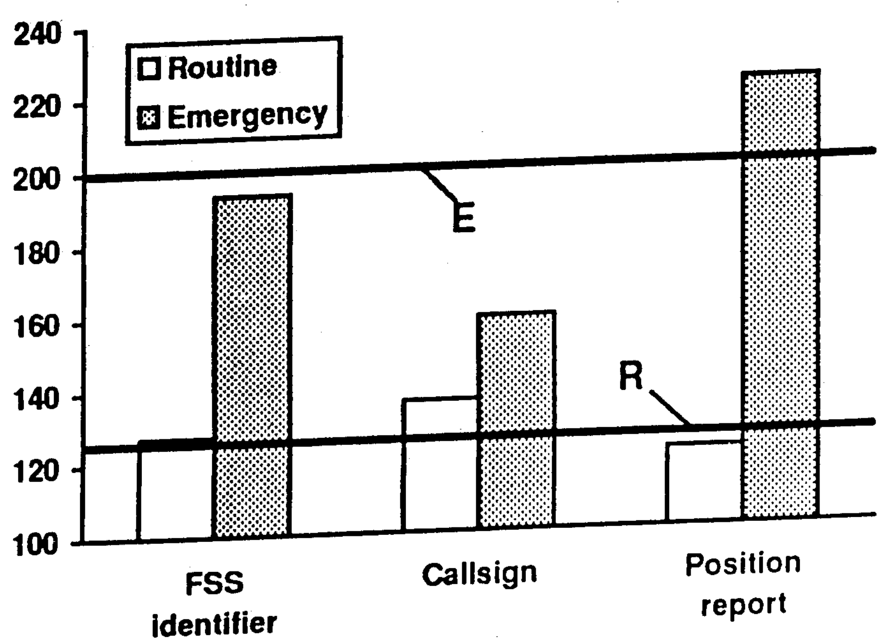


FIGURE 2: Fundamental Frequencies of Phrase Pairs, and Mean Fundamental Frequencies of all Routine (R) and Emergency (E) Statements

Further, as Belan predicted, the range of fundamental frequencies within statements grew larger under emergency conditions, and the number of syllables per statement decreased. The real value of this technique will lie in its ability to determine information about the emotional state of a speaker when it is not otherwise apparent. It is hoped that the technique described in this paper will lead to the ability to do just that in a standardized way. A tool for exploring the cognitive and emotional states of people involved in accidents could prove invaluable in determining the underlying causes of their performance and identifying appropriate preventative strategies.

Author's Note: A version of this paper was published in the Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting in October 1994. All opinions expressed in this paper are those of the authors and do not necessarily reflect the official position of the National Transportation Safety Board.

REFERENCES

Belan, A.S. (1994). Personal communication between the authors and Alfred S. Belan MD, Chief of the Acoustic Laboratory, Interstate Aviation Committee, Moscow, during his visit to the U.S. National Transportation Safety Board, Washington, D.C., February 14-18, 1994. [Dr. Belan has published his work in Russian, but it is not currently available in English.]

Brenner, M., and Cash, J.R. (1991). Speech analysis as an index of alcohol intoxication: The EXXON VALDEZ accident. *Aviation, Space, and Environmental Medicine*, 62, 893-898.

Brenner, M., Doherty, E.T., & Shipp, T. (1994). Speech measures indicating workload demand. *Aviation, Space, and Environmental Medicine*, 65, 21-26.

Brenner, M., Shipp, T.E., Doherty, T., & Morrissey, P. (1985). Voice measures of psychological stress: Laboratory and field data. In I.R. Titze and R.C. Scherer (Eds.), *Vocal fold physiology, biomechanics, acoustics, and phonatory control* (pp. 239-248). Denver: The Denver Center for the Performing Arts.

National Highway Traffic Safety Administration. (1989). Briefing paper: *Drug evaluation and classification program.* Washington, DC: U.S. Department of Transportation.

National Transportation Safety Board. (1990). *Marine accident report: Grounding of the U.S. Tankship EXXON VALDEZ on Bligh Reef, Prince William Sound, near Valdez, Alaska.* March 24, 1989 (NTSB Report No. NTSB/MAR-90/04). Washington, DC: National Transportation Safety Board. (NTIS No. PB90-916405).

Ruiz, R., Legros, C., & Guell, A. (1990). Voice analysis to predict the psychological or physical state of a speaker. *Aviation, Space and Environmental Medicine*, 61, 266-271.

Scherer, K.R. (1981). Vocal indicators of stress. In J.K. Darby (Ed.) *Speech evaluation in psychiatry.* New York: Grune and Stratton.

Streeter, L.A., McDonald, N.H., Apple, W., Krauss, R.M., & Galotti, K.M. (1983). Acoustic and perceptual indicators of emotional stress. *Journal of the Acoustical Society of America*, 73, 1354-1360.

USING OCS TOOLS™ IN TEAM PERFORMANCE RESEARCH

Clint Bowers, Florian Jentsch, Barbara Holmes
University of Central Florida

INTRODUCTION

Analyzing communications of team members has become an important method in the area of team performance research (cf. Bowers, Braun, & Kline, 1993). Analyses of intra-team communications allow an outside observer one of the few opportunities to gain an understanding of the cognitive and social processes occurring within teams. However, after team performance data are collected and an adequate coding scheme has been developed, 2 problems are encountered in communications coding: The selection of the hardware and software to perform the coding and the actual coding procedure. Possible approaches include paper- and pencil-based coding and manual data entry, or computerized coding, data entry, and analysis. The Team Performance Laboratory uses both manual and computerized methods, depending upon the scope of the analyses and the available data. For the computerized analyses, the Team Performance Laboratory employs OCS TOOLS™, a software and hardware package developed by Triangle Research Collaborative, Inc. OCS TOOLS™ was selected by the Team Performance Laboratory because we needed a data analysis tool that was flexible enough to be useful in a variety of research applications. Our main focus was on communications analysis, but we also wanted to perform network analyses, tactical decision-making analyses, and task analyses. The OCS TOOLS™ system answered this statement of needs because it allows for the coding and simultaneous timing of live or videotaped events according to a variety of coding schemes. Using the system, timestamped videotapes are coded by a trained rater on a basic workstation consisting of a personal computer with monitor and keyboard connected to a video cassette recorder (VCR). The output datafiles provided by OCS TOOLS™ are ready for further statistical analyses using standard statistical software packages. In the following sections, we describe these problems in more detail. For each problem, we also show how the Team Performance Laboratory has implemented a solution, and what experiences we made with these solutions.

What are the Tools for Coding? - Description of OCS TOOLS™

The Observational Coding System (OCS TOOLS™) by Triangle Research Collaborative, Inc. is an integrated software and hardware system for observational data collection, preliminary data analysis, and records management. In the Team Performance Laboratory, OCS TOOLS™ are mainly used for coding of intra-team communications and crew coordination behaviors. The system allows the researcher to combine observational methods with computer and video technology into an integrated whole. This can increase the reliability of the codings and often allows for easier data storage and handling than traditional manual coding systems.

Basic Architecture

OCS TOOLS™ consist of several hardware and software modules which can be assembled in a variety of architectures. Three basic systems, called LIVE, FRAME, and VCR, allow customization of the OCS TOOLS™ set to a variety of research settings. With OCS-LIVE, events are coded as they occur by entering the appropriate code (TRC, 1993). OCS-FRAME, on the other hand, includes the features of OCS-LIVE, but also allows the coder to enter a user-selected time code with each code. Finally, OCS-VCR can perform the functions of OCS-LIVE and OCS-FRAME. Alternatively, OCS-VCR can use a machine-readable timestamp from the videotape that is coded as a timing reference.

Hardware Components. Several hardware modules make up the OCS TOOLS™ system. A timecode reader reads optional timestamps from the audiotracks of a videotape and automatically records time in the data stream. Also, a VCR controller allows the optional control of a VCR from the keyboard of the OCS TOOLS™ computer. A second keyboard can be connected to the system, allowing 2 coders to rate the same event simultaneously. Further options that are available include the capturing of keystrokes from an independent computer (for the purposes of software usability testing) and video overlaying. The latter option allows viewing of VCR and computer interface simultaneously or multiplexed on the same monitor.

Software Components. All systems have several common features. They share functions for basic statistics (frequency and durations of specified events, analyses of intervals between events, time series comparisons, and pattern analyses). All OCS TOOLS™ systems also have a common package of software utilities. These routines allow operators to manipulate files, gain access to directories, etc. Also, each system contains advanced functions, ADMIN, AGREE, and PLAYBACK. The ADMIN functions allow the selection of hardware and software components to be used for a particular coding task.

Furthermore, they allow a system administrator to monitor the progress of coding through an audit trail, to limit access to files, and to specify other variables related to data security. PLAYBACK and AGREE can be used for dataset verification and observer training. With PLAYBACK, the operator can review data sets to specified points. Also, this utility allows identification of trouble spots by presenting the codes and the respective videorecording simultaneously. AGREE, on the other hand, allows the researcher to compare 2 sets of data to verify interrater reliability and code consistency. The input and output files are all in ASCII format and are therefore compatible with many DOS-based software packages.

Current System Layout in the Team Performance Laboratory

The Team Performance Laboratory uses a single computer, monitor, VCR, and keyboard in its OCS-VCR configuration. The single computer is an IBM-compatible 80286-personal computer that is connected to a professional VCR. This setup is sufficient for the purposes of the Team Performance Laboratory, as it allows laboratory staff to play videotapes of aircrews engaged in complex flight scenarios and code their interactions in real-time.

How Do We Code? - Practical Applications. Coding behaviors as they occur involves significant problems: Obtrusiveness of the raters, reactivity from the participants to the presence of raters, the limited capacity of raters to remember and rate communications, lost time if raters are present at a site without observable events, etc. reduce the effectiveness of the rating process. Furthermore, it is difficult to keep raters unaware of the treatment condition ("condition-blind") if they are present at the observation site. Because of these problems, the Team Performance Laboratory has selected to video- and audiotape the interactions within the experimental teams and to rate these recordings after the fact in a laboratory. While this method introduces its own set of problems (e.g., identifying speakers from audiotapes), it allows the events to be rated in a randomized order and helps coders to remain "condition-blind."

Within the methods that use recordings as the basis for coding of team communications, videotapes are preferable over audiotapes. The Team Performance Laboratory has equipment to timestamp videotapes (see below), but not for the timestamping of audiotapes. Also, video tapes can facilitate the identification of the speakers, provided their pictures are recorded. We found in the Team Performance Laboratory that raters have particular difficulties distinguishing among the voices of pilots. Causes contributing to these problems are that most participants in flight simulations are male, of about the same age, and come from a relatively limited geographical area within the U.S. Furthermore, the headset and microphones used by the participants, while increasing the physical and functional fidelity of the simulation, often do not provide optimal transfer characteristics for audiorecordings; another factor making the identification of speakers from audio recordings alone very difficult.

Another advantage of using videotaped communications is that the video often helps coders to classify otherwise ambiguous communications. The visual information about who is manipulating the controls, which chart a pilot is looking at, or which instruments he/she is pointing at, can be very useful when categorizing communications.

Timestamping. The OCS TOOLS™ software allows a computer system to function as an event recorder, which the rater uses to code and record events as they occur. Data from coding sessions are stored directly to disk and may be edited later. This way, events may be coded live in the field, or videotaped and coded later. When events are coded live, each time a code is entered at the keyboard, it is assigned a time using the computer's internal clock. The code and the time it was entered are saved in the dataset.

Coding live is often impractical or impossible for research purposes: The amount of data that needs to be processed, evaluated, translated into a code, and physically entered into the computer may quickly exceed the capabilities of even the best trained coder. As a result, even a well-trained coder may miss events that need to be coded. Therefore, the Team Performance Laboratory makes use of the other coding option for OCS TOOLS™, that of using pre-recorded videotapes. When events are coded using this method, the OCS TOOLS™ system can operate either in a synchronous or non-synchronous mode with the VCR. Non-synchronized means in this context that the computer uses its internal clock to assign a time each time a code is entered. Although videotape can be coded this way, it is not done in the team Performance Laboratory because of the coders' limitations discussed above that may make the time assigned to each coding unreliable.

Rather than coding videotapes in the non-synchronous mode, the Team Performance Laboratory uses the synchronized mode. Using this method, each videotape is timestamped before it is coded; that is, each frame of the videotape is stamped with a time code that the computer can read. When a code is entered, the computer assigns it the timecode read from the respective frame of the videotape. Even if the videotape is rewound or fast forwarded, the computer enters the correct "video timestamp." Thus, unlike in the non-synchronous mode, coders can rewind, recode, fast forward, or code at any tape speed the system can accommodate, without worrying about incorrect times being assigned to codes of events.

The first step of preparing a videotape for coding is therefore to lay a timestamp on the tape which can be read by the OCS TOOLS™ system. The timestamp may be laid onto the tape at the time of recording, or it may be copied onto a duplicate tape. Copying a timestamp onto a duplicate tape is time consuming since timestamping must occur at the original tape speed (i.e., high-speed dubbing cannot be used). The Team Performance Laboratory therefore uses a special timestamp generator at the time of the original recording. This minimizes the delay between data collection and coding of data.

What Did We Experience? - Lessons Learned and Outlook

From its use in the Team Performance Laboratory, we have learned several important lessons about OCS TOOLS™ and their utility for the coding of intra-team communications. The following is a compilation of some of the advantages and disadvantages that we found in our experience with the coding of communications using OCS TOOLS™.

Advantages

Computing Power. As OCS TOOLS™ is DOS-based, it can be run on any IBM-compatible processor (AT or better). Thus, it can be run on a relatively inexpensive PC, reducing the equipment cost required. OCS TOOLS™ can also create, edit, and store datafiles of various sizes and complexity. The user is only limited by the memory capacity of the computer OCS TOOLS™ is run on. Also, its output of ASCII files can easily be read by most conventional statistical packages such as SPSS and BMDP.

Flexibility. One of the most flexible aspects of OCS TOOLS™ concerns the assignment of codes. Users of the system are free to design any types of coding scheme they desire, with the maximum limit being 10 characters. This allows researchers to pick and choose the most appropriate coding scheme suited for their use. The editing feature even allows investigators the flexibility of altering datasets subsequent to their creation. Should a coding scheme be redesigned after coding of participant interactions has begun, researchers can change the previous datafiles to adhere to the newer coding scheme. This editing feature is generally representative of the entire system's flexibility.

Customizing. OCS TOOLS™ can easily be configured according to the needs of the user. One is generally limited only by the amount and type of hardware available. Should a particular configuration not be diagrammed in the instructions, TRC staff are willing and able to help system operators to design optimal configurations for their research needs.

Disadvantages

Interface. The main disadvantage we found while using the OCS TOOLS™ system in the Team Performance Laboratory is that the interface of the system is less intuitive than we expected. We found that observers require a thorough training session before we can confidently let them use the system. Our lab employs a large number of undergraduate students that conduct directed research for only one to two semesters. Before these undergraduate students can work as observers under the supervision of subject matter experts, they have to be trained in using the coding schemes and with respect to the subject matter. Training prospective raters to use OCS TOOLS™ imposes additional demands on the subject matter experts, and often is not justified if the raters are going to work in the laboratory for only a few months. In fact, at this time, we are not training new raters to use OCS TOOLS™ because of this problem.

Code Limits. Even though a 10-character limit would not seem detrimental, when coding in real time it is often difficult or even impossible to type in 10-character codes when interactions are occurring rapidly. Coders cannot possibly keep up with their observations because the quantity of characters soon exceed the capacity of their working memory. It is therefore advisable for users of OCS TOOLS™ to limit the number of characters used in their coding schemes to as few as possible in order to expedite the coding process. We in the Team Performance Lab generally utilize 2-character codes to identify not only the speaker but also the type of statement uttered.

Timing and Recording. The timestamping of videotapes via OCS TOOLS™ is time-consuming and tedious. It requires the use of 2 VCRs connected to the computer using OCS TOOLS™, as well as an alternate wiring scheme than that used when coding tapes. The switching back and forth between wiring configurations can lead to errors, and therefore, annoyance. Additionally, when OCS TOOLS™ is in the timestamping setup, it cannot be used to code tapes at the same time. Thus, timestamping videotapes reduces the amount of the time the computer can be used to code datasets. To reduce the additional time required to timestamp via OCS TOOLS™. Team Performance Lab staff have resorted to using another timestamp method which does not require the use of OCS TOOLS™, and also allows videotapes to be timestamped during the original recording. It also increases the amount of time the OCS TOOLS™ computer can be used by coders.

OUTLOOK

As can be seen from the previous discussion, not all coding tasks within the Team Performance Laboratory are completed using OCS TOOLS™. In fact, in many cases, we found that it is faster to have raters code certain behaviors manually. This is especially the case if ratings have to be made quickly, and in cases where sequences of events are not as important as the frequency of their occurrence. In these cases, raters can code videotaped events at any VCR without the need of using the specific OCS TOOLS™ workstation. This reduces the time required to have multiple raters code videotapes. Also, this method often increases the acceptability of the coding process with raters who are not confronted with the logistical problems of sharing a workstation at a particular location. We therefore decide about the use of OCS TOOLS™ on a case-by-case basis, rather than always using the system.

In those cases that the Team Performance Laboratory has used OCS TOOLS™, it was only employed in a limited capacity. This is in part the result of the limited hardware set present in the Team Performance Laboratory (1 computer and 1 VCR), but was also partly based on the fact that we did not need all the functions offered by the system. One such function that is currently not used by the Team Performance Laboratory but may be utilized in the future is the simultaneous coding of 1 videotape by multiple raters at individual workstations. This approach has special utility when the same events are to be coded using different coding schemes, or if raters are focusing on different persons, objects, or behaviors (e.g., one rater rates verbal communications, the other codes non-verbal signs). Also, concurrent coding by several raters can be used to perform rater training more effectively, and to quickly establish the degree of interrater reliability.

As is shown by this example of a future application, OCS TOOLS™ provide a large number of functions that are limited mainly by the financial resources available to the user. As future tasks will impose new requirements for communications coding and analysis, the Team Performance Laboratory will expand the use of this tool to fulfill these needs.

REFERENCES

Bowers, C. A., Braun, C. C., & Kline, P. B. (1994). Communication and Team Situational Awareness. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 305-311). Daytona Beach, FL: Embry-Riddle Aeronautical University.

TRC (Anon.). (1993). *OCS TOOLS - Observational coding system of tools* [Product description]. Research Triangle Park, NC: Triangle Research Collaborative.

AUTHOR NOTES

OCS TOOLS™ is a registered trademark of Triangle Research Collaborative, Inc. (TRC), P.O. Box 12167, 100 Park, Suite 115, Research Triangle Park, N.C. 27709.

The Team Performance Laboratory and its staff are not affiliated or associated with Triangle Research Collaborative, Inc. All claims and representations made in this paper about the capabilities of OCS TOOLS™ are based on actual experiences made with the system at the Team Performance Laboratory of the University of Central Florida and may therefore not be completely indicative of the system's capabilities in other settings.

Correspondence concerning this article should be addressed to Dr. Clint Bowers, Department of Psychology, University of Central Florida, Orlando, Florida 32816-1390, internet: bowers@pegasus.cc.ucf.edu.

ANALYZING VOICE COMMUNICATION CONTENT: WHAT MacSHAPA OFFERS

Penelope M. Sanderson
Department of Mechanical and Industrial Engineering
University of Illinois at Urbana-Champaign

INTRODUCTION

In many domains of inquiry we need effective ways of analyzing human verbal and non-verbal communication. The analysis of verbal communication has traditionally been supported with audiotape and transcription, but is now increasingly supported with videotape. The analysis of non-verbal communication, however, relies heavily upon videotape to provide a record of gesture, expression, bodily orientation, and direction of gaze in addition to verbal information. When human communication is studied in high technology working environments such as aviation, process control, or hospital operating theatres, data collection becomes even more complex. If human communication is to be understood in such work contexts, important features of the context need to be tracked and stored alongside the verbal and non-verbal communications data.

For example, when studying communication patterns in a cockpit simulator, we might collect several video signals (from 2 or more video cameras positioned at different locations), an electronic log of crew actions, information about aircraft status sampled many times per second for many parameters, and finally environmental information such as wind direction, outside temperature, etc. To recapture the work context and fully understand what the human participants were achieving as they communicated, we must be able to coordinate these different sources of data so that their interrelations are apparent.

Coordinating such data is difficult both technically and conceptually. Over the last 5 years there has been considerable progress in surmounting some of the technical problems (see review in Sanderson, 1994). The arrival of relatively low-cost multimedia hardware and software has encouraged many researchers to build data analysis environments that are equal to the challenge of rich communication data. Less progress has been made on the conceptual front, however, largely because the overwhelming task of first gaining access to the data still leaves us with less time to explore the data and try out different forms of analysis than we would like. Therefore, investigators still face dilemmas on all fronts when deciding how to analyze communication data (Sanderson & Fisher, 1994). For example, what aspects of the data should be highlighted, how should the data be sampled if all of it cannot be analyzed, should data be "coded" or loosely described, what kinds of statistics, if any, can be used, and what constitutes adequate "proof" of an assertion?

Answers to these questions depend partly on the intellectual tradition to which an investigator belongs (such as ethological, cognitive, interactionist, ethnomethodological, etc.). However, answers about how to proceed also depend greatly on the question that is being answered with the data, and on the form of the data themselves. There has been a flurry of writing about the connection between technical and conceptual aspects over the last few years that has helped make us more aware of the choices to be made and the basis on which they can be made (Edwards & Lampert, 1993; Fielding & Lee, 1991; Jordan & Henderson, in press; Sanderson, 1993; Sanderson, 1994; Sanderson & Fisher, 1994; Weitzman & Miles, 1994).

In this paper I will briefly describe a program called MacSHAPA that has been under development at University of Illinois for the past 4 years. MacSHAPA is a Macintosh-based application with simple multimedia capabilities that helps the analysis of certain kinds of sequential data, including verbal and non-verbal communications data. MacSHAPA was initially developed to help analyze cockpit communication, but because it is a "context-free" tool it can be applied equally well to the analysis of observational or sequential data in many different domains.

MacSHAPA's Structure

MacSHAPA's structure can most easily be described with the "star diagram" in Figure 1 and the interface example in Figure 2. MacSHAPA's basic data display is a special kind of spreadsheet, as Figure 2 shows. The columns (which we call "variables") hold different kinds of data such as transcription, a researcher's notes, electronically captured control activity, etc. Within each column are small boxes, which we call "cells." Cells hold the elements of information in each column, such as a single utterance, a single action, etc. Variables and cells are at the heart of a MacSHAPA document, so they have been placed at the center of the star diagram in Figure 1.

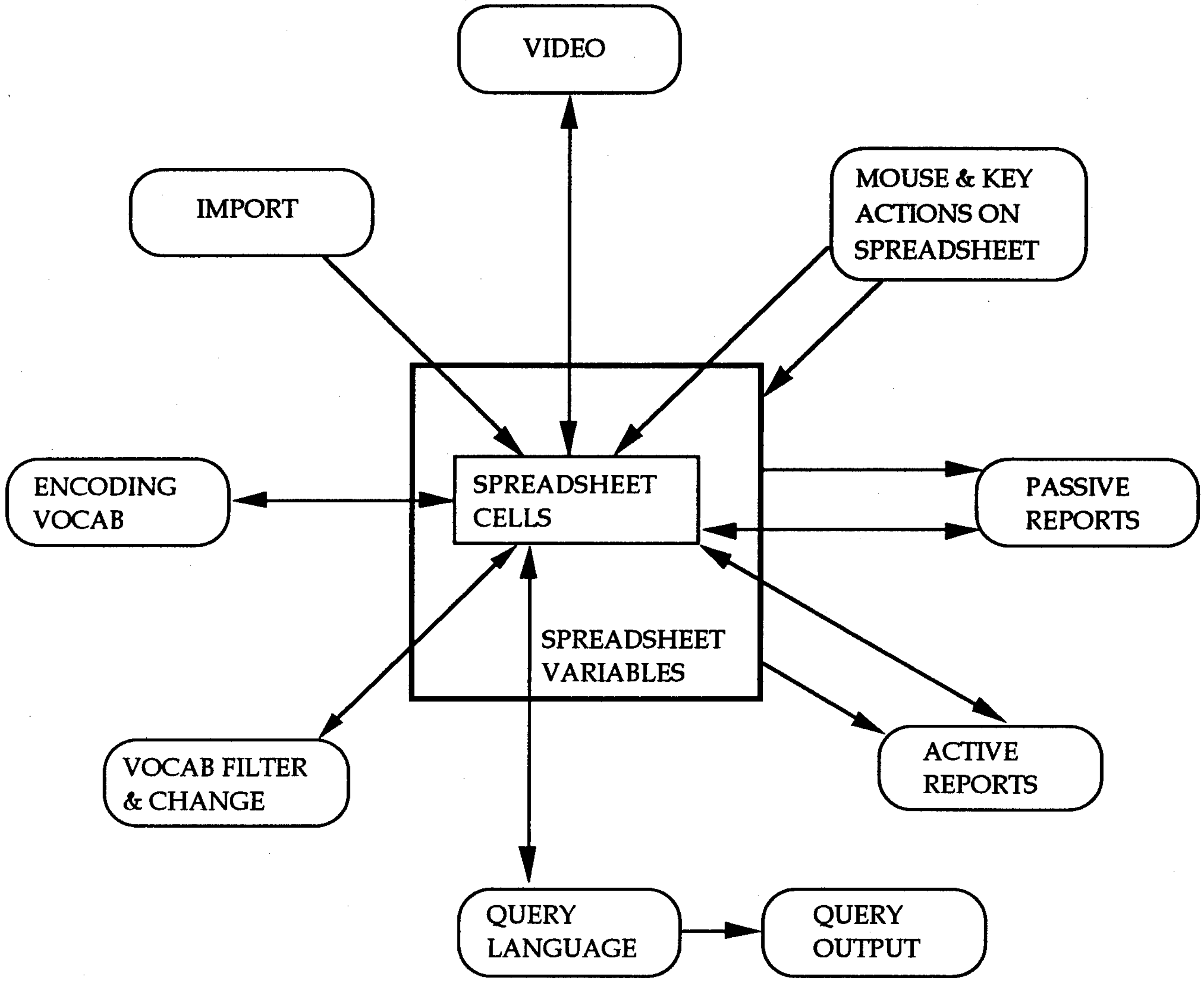


FIGURE 1: Star Diagram of MacSHAPA's Functionality.
In center, "spreadsheet variables" indicates columns of different kinds of data, and "spreadsheet cells" represents the data atoms or elements within each variable, or column.

Around the perimeter of the star diagram are the names of the most important functions people carry out with MacSHAPA—making mouse and key actions on the spreadsheet to enter and edit raw data, handling video, importing data from other applications, setting up encoding vocabularies (or coding schemes), filtering and changing encoding vocabularies, and formulating queries in a general query language. The functions break down into 3 general classes of activity, which will be discussed in greater detail in the next 3 sections.

1. Seeing data in various ways (includes video, mouse and key actions, passive reports).
2. Entering and editing data (includes mouse and key actions, video, import, encoding vocab, vocab filter, and the query language).
3. Carrying out analyses and statistical procedures on data (includes query language, active reports, passive reports).

1. Seeing Data. Figure 2 illustrates some of the many ways that data can be seen in MacSHAPA: in video form; in spreadsheet form as transcriptions, encodings, or annotations; and in a visual timeline representation (lower right). The data in the spreadsheet can include transcriptions, comments, encodings, and theoretical annotations.

If video is being used, then MacSHAPA provides remote control of a video source through a VCR Control window that includes all normal VCR commands, plus some further useful commands (see top left of Figure 2). The VCR Control window lets the user see data in the following ways:

- Control the basic movements of the videotape such as play, pause, stop, forward, and rewind.
- Control the jog and shuttle functions.
- Search for a specified timecode on the videotape.
- Replay video and see data cells in the spreadsheet highlight in synchrony with the videotape, as their timestamps match the timecode on the videotape.

MacSHAPA has a built-in driver that controls a Panasonic AG-7750 VHS/ SVHS professional level VCR with an onboard AG-F700 timecode generator/ reader card. This driver also works with Panasonic's newer VCR models. MacSHAPA can also control various other video devices with the help of Abbate Video Inc.'s VTK Remote™ application. Using Apple Computer Inc.'s Video Monitor™ the video signal can be digitized and sent to the computer screen, as seen in Figure 2.

Users can select data in the spreadsheet using standard mouse and key actions, and then ask to see the data in different forms (the so-called "passive" reports in the star diagram). For example, selected variables (columns) or cells can be viewed in a graphical timeline, as shown in Figure 2, or in a more compact listing form rather than as a spreadsheet. Active links are maintained between data in the spreadsheet, positions on a timecoded videotape, and graphical representations of events in a timeline display.

The layout of the spreadsheet itself can be changed. The first timestamp in each spreadsheet cell is the cell's starting time and the second timestamp is its ending time. In Figure 3a, cells are positioned so as to preserved a weak temporal ordering in the timestamps across different columns, and the timestamps are displayed. In Figure 3b, however, the cells have not been positioned to preserve weak temporal ordering, but instead just to save space. Additionally, the timestamps for each cell are not drawn. This leads to a more compact representation, especially if a column is narrowed as well (not shown here).

2. Entering, Editing, and Manipulating Data. The nodes labeled "Mouse and key actions on spreadsheet," "Video," and "Import" all contribute to entering data into MacSHAPA. Through mouse and key actions, users can perform many functions directly on the spreadsheet representation of the data. These functions include creating new data columns, entering new data cells into the columns, and changing the look and layout of the spreadsheet by moving columns and cells from place to place.

As we have seen, users can control a VCR remotely through MacSHAPA's VCR Control window. In addition, users can capture timecodes from a video source that has timecode stored on it or from the Macintosh's internal clock, and insert timecodes into spreadsheet cells. This process is illustrated in Figure 4. While the VCR time counter or the internal clock runs, users create new cells by hitting the Stamp New Cell button on the VCR Control panel. A new cell will be created and the time of its creation will be automatically inserted into its time onset. The user can then enter a comment or code.

With the help of QuicKeys®, the user can create "coded event buttons." A coded event button is a key that, when pressed, creates a timestamped new cell and inserts a code or description into the cell, such as "Redirects Captain's attention non-verbally" or "Raises voice." Clearly, a well-conceived set of coded event buttons can save a great deal of time-consuming typing and allow quite complex coding and annotation to take place in real time, such as when observing events in the field or working with videotape.

The "Import" node in the star diagram (see Figure 1) refers to the fact that users can import external text files into MacSHAPA. For most importing needs, MacSHAPA's general format configuration will be adequate. Users tell MacSHAPA what the structure is of each record in the raw data file is and where the data should go in MacSHAPA's spreadsheet, and MacSHAPA will do the rest.

For some research needs, but not all, it helps to develop a strict coding scheme to apply to the data. The node "Encoding vocab" indicates that users can set up templates or vocabularies for encoding. As the node "Vocab filter" suggests, users can filter their data, selecting some parts and ignoring others, and then either perform reports on the filtered data or rewrite the filtered data in some way. Filtering and changing data helps new ways evolve of describing and understanding data.

Finally, the node labeled "Query language" refers to MacSHAPA's database query and data manipulation language. Each query consists of a condition and an action. In the condition the user defines a certain pattern to be sought in the data, such as "first utterance after turbulence encountered." In the action the user states what should happen whenever that pattern is found. The query language can be used for inserting new cells, modifying old cells, deleting cells, and selecting certain cells for further analysis.

3. Analyzing and Reporting Data. Before running a MacSHAPA report, users must identify the data on which the report should be run. As Figure 5 shows, queries and reports can be run on spreadsheet selections of data (cells or variables), or on data sets created by a filtering operation. Selections can also

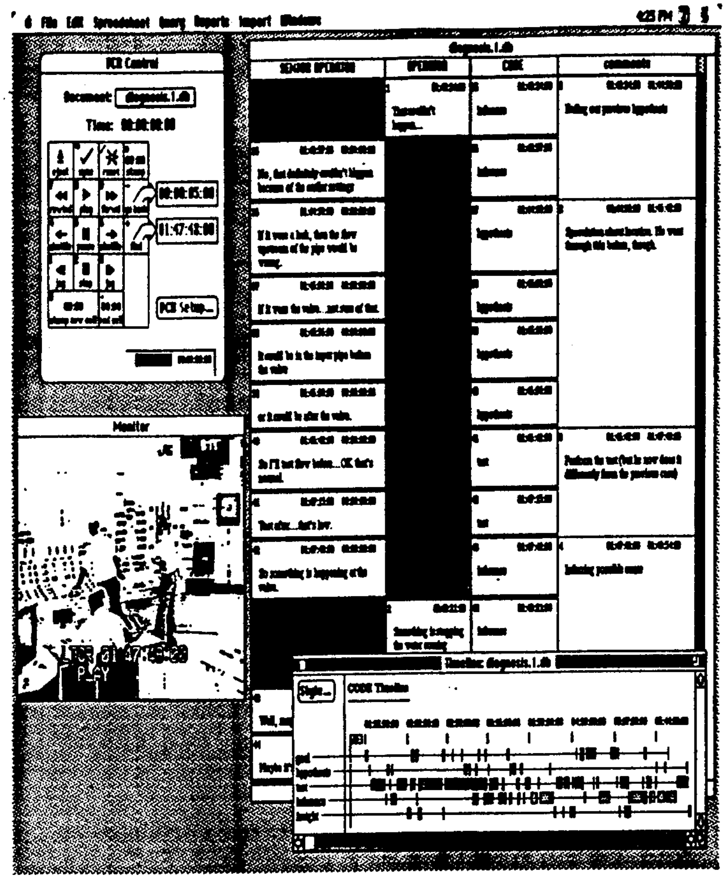


FIGURE 2: Different Views into Data with MacSHAPA.
Dominating the right background is the spreadsheet-like data display, its columns containing qualitatively different kinds of information. At bottom right is a timeline display of the codes in the "CODE" spreadsheet column. Digitized video is shown bottom left, and the VCR Control window at top left.

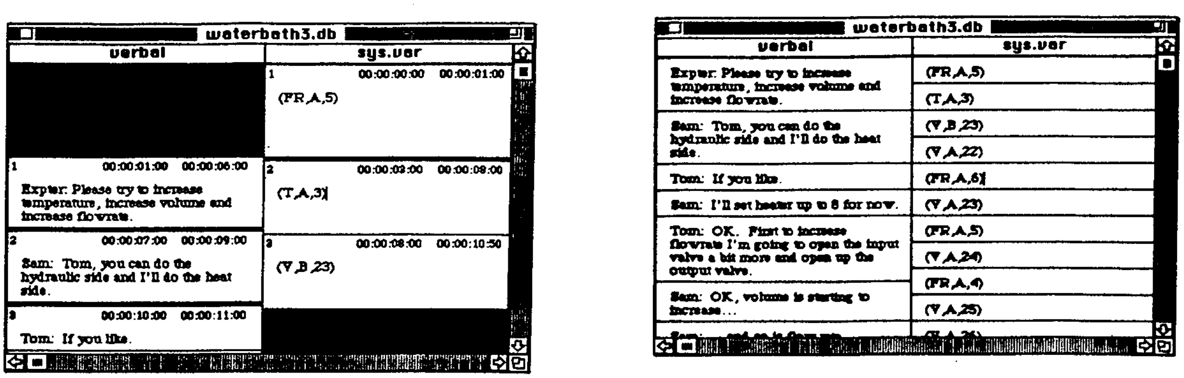


FIGURE 3: Alternative Spreadsheet Layouts.
(a) Temporal ordering on and timestamps drawn, (b) temporal ordering off and timestamps not drawn.

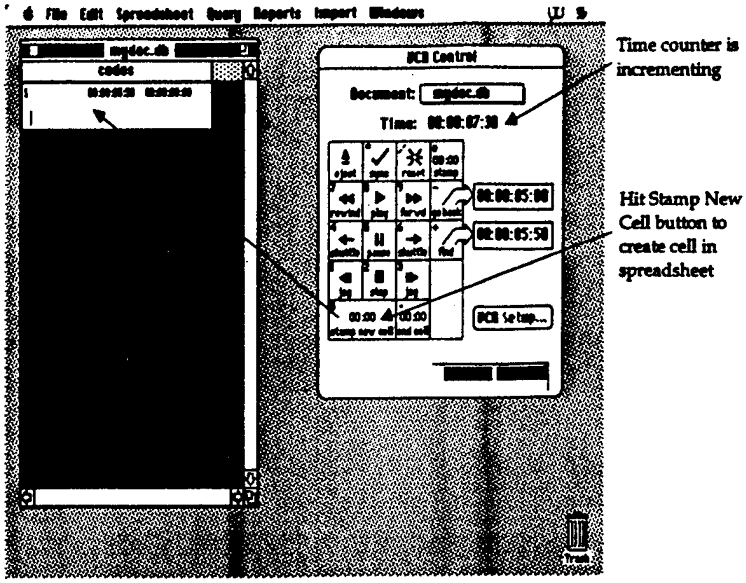


FIGURE 4: Using the VCR Control window's Stamp New Cell button, users can capture timecode from an external source and create a new cell in the spreadsheet with the captured timecode in its time onset. (Time readout in VCR Control window is later than time in cell because picture was taken about two seconds after cell was created.)

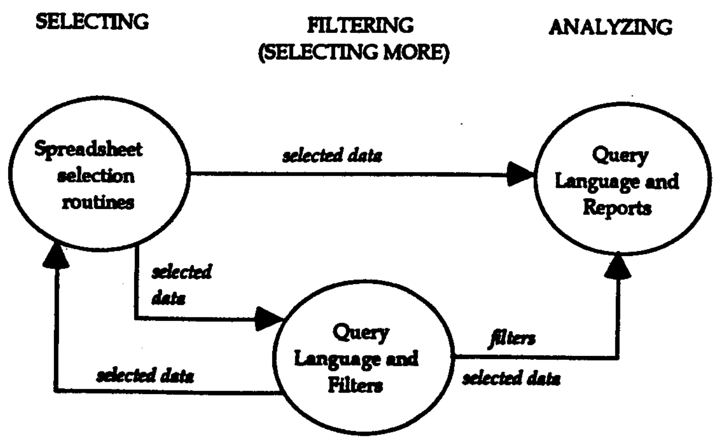


FIGURE 5: In MacSHAPA, queries and reports can be run on selected spreadhseet cells, selected spreadsheet variables, or on data created by a filtering operation.

be modified by further selecting, querying or filtering, narrowing the data chosen until just the desired subset is selected to go forward to a report.

There are 2 principal ways of analyzing and reporting data in MacSHAPA—using built-in reports and using the query language. Reports can be *passive* and *active*. Passive reports are run simply by selecting 1 or more columns of data or a set of individual data cells on the spreadsheet, and choosing a passive report in MacSHAPA's Report menu. Passive reports include timeline analysis which helps to detect patterns (see Figure 2), content analysis which itemizes and counts how different codes are used, and duration analysis which reports how long each code was active.

In contrast, active reports require some settings and selections to be made in a dialog box before they can be run. They include transition matrices with some simple Markov statistics, analysis of cycles between key events, lag sequential analysis, and the comparison of different event streams with either reliability measures, information transition measures, or a basic time-warping routine. Further details can be found in Sanderson et al. (in press).

Figure 6 shows a MacSHAPA document (left) with the results of an active report (transitions analysis) and a passive report (content and duration analysis) next to it. Both reports have been performed on the "CODE" column. The transitions analysis has been printed out as a tree, although the more conventional matrix can be generated. The data suggest that in this fault diagnosis episode, statements about tests are very common and are usually followed by inferences. The content analysis reports, for each code, the number of times it was encountered in the document, the total amount of time the code was active, and the average time (per occasion) that it was active.

Finally, the Query language can be used for further types of reports and analyses. It can be used to count events, sum values, perform arithmetic and Boolean operations on cell values, and search for simple sequential patterns. The basic query template is:

query(<condition>,<action>)

In the <condition> side the user enters patterns to seek in the data, and in the <action> side enters what should be done when the patterns are found, such as printing them out, selecting the cells found, adding 2 minutes to their timestamps, etc.

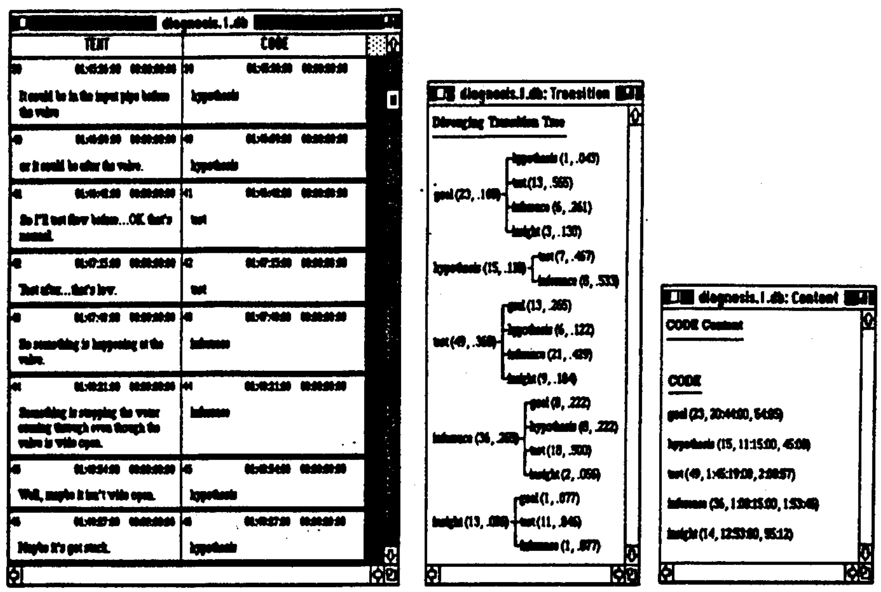


FIGURE 6: At left, a coded transcript. At center, transition analysis of these data displayed as a diverging tree. At right content and duration analysis on the same data.

Figure 7(a) shows 2 queries in a MacSHAPA document that uses a complex relational template for encoding, in which there is a key term ("ACKNOWL") followed by some qualifiers: ACKNOWL (<SPEAKER>, <TO>, <MITIGATION>). The first query looks at cells in the column called "speechcode" and finds all cells in which Tom makes an acknowledgment—ACKNOWL(Tom, <TO>, <MITIGATION>). The action is then to count the number of times acknowledgments by Tom are found, and the result (Count=4) is shown in Figure 7(b).

The second query again looks for acknowledgments by Tom, and stores the ordinal number of the cell in "?ord" and the time onset of the cell in "?on." The action is to print out the number and time onsets of cells in which Tom makes an acknowledgment.

Fortunately, users do not have to type in all the punctuation shown in the sample queries above in Figure 7. The query language has a structure editor that "explodes" with the proper syntax and manages the punctuation in the background. The query language is for advanced use—many of the analyses that can be performed with it can also be performed more simply with reports, but with the query language users can pose unusual questions and carry out very specific transformations.

Suitable Uses of MacSHAPA

Some human communication investigations can be conveniently supported with MacSHAPA, whereas it is less suitable for other kinds. For example, MacSHAPA was designed to be used primarily with

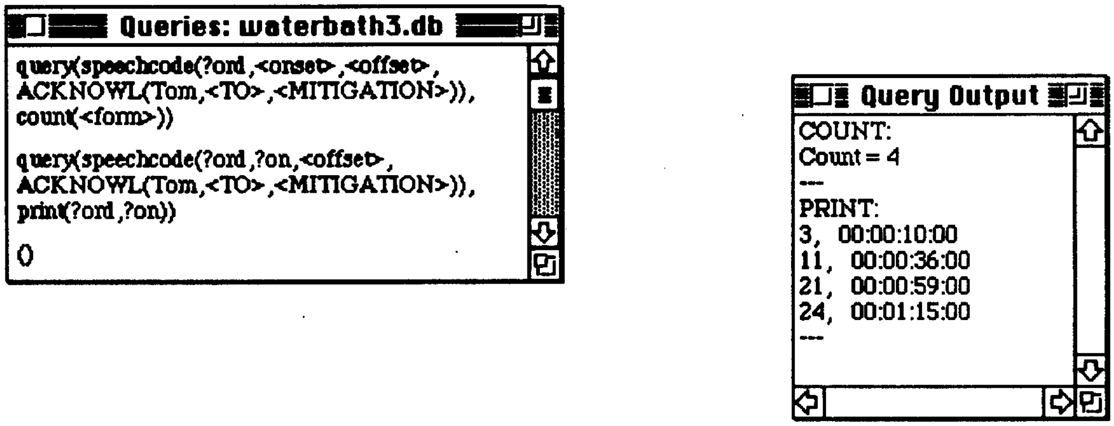


FIGURE 7: Sample Use of the Query Language
(a) Queries window with two queries. (b) Output of each query in the Query Output Window

symbolic data, such as codes describing human and system activity. At present, MacSHAPA has less to offer to the analysis of strictly numerical data such as a speech signal or raw eye movement data. MacSHAPA helps investigators develop and change coding categories, store them, and use them to encode data manually. There are no coding categories "built into" MacSHAPA; the software does not encode data automatically.

Temporal relations are an important organizing principle in MacSHAPA, which makes it suitable for analyzing temporal aspects of communication. Comments and annotations, as well as events, are associated with a particular point in time. Because of this, MacSHAPA is particularly useful for analyzing sequential and linear aspects of observational data but is of less help when analyzing nonlinear aspects.

Examples of Use in Verbal and Non-verbal Communication Studies

MacSHAPA has now been used in several investigations involving human verbal and non-verbal communication. It has been used by the Aeronautical and Maritime Research Laboratory in Australia to analyze audio tapes of intercom communication between crewmembers on several P3-C Orion surveillance aircraft during full-scale exercises (Manton, personal communication, 1992). It has been used to study collaborative reasoning in scientific discussions (Dunbar & Baker, 1993) and diagnostic reasoning (Reising, 1992). In addition, MacSHAPA has been used at NASA Ames Research Center to study voice communications in party line and data link ATC configurations (Mosier, personal communication, 1994) and non-verbal communication between aircraft crewmembers (Segal, 1993).

Obtaining MacSHAPA

MacSHAPA represents the implementation of a preliminary hypothesis about how certain kinds of ESDA might be aided. The software is primarily a research tool developed in a research laboratory, and does not have some of the features expected of a commercial software product. However, it is continually evolving in response to user comments. Copies and upgrades of MacSHAPA can be obtained from CSERIAC at Wright-Patterson Air Force Base. For more information about MacSHAPA and to obtain a copy of MacSHAPA, contact CSERIAC through one of the following routes.

{CSERIAC Program Office
AL/CFH/CSERIAC Building 248, 2255 H Street
Wright-Patterson AFB OH 45433-7022
Tel: 1 (513) 255-4842}

Alternatively, you can contact CSERIAC's technical transfer specialist, Mr. Chris Sharbaugh, at: csharbaugh@falcon.aamrl.wpafb.af.mil

NOTES

QuicKeys® is a registered trademark of CE Software, Inc. Video ToolKit™ is a trademark of Abbate Video, Inc. Video Monitor™ is a trademark of Apple Computer, Inc.

REFERENCES

Dunbar, K. & Baker, L. (1993). *Group problem solving: What is it and when does it succeed?* Technical report, McGill Cognitive Science Center, Department of Psychology, McGill University, Montreal, Canada.

Edwards, J.A. & Lampert, M.D. (1993). *Talking data: Transcription and coding in discourse research.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Fielding, N.G. & Lee, R.M. (1991). *Using computers in qualitative research.* London: Sage.

Jordan, B. & Henderson, A. (In press). Interaction analysis: Foundations and practice. *Journal of the Learning Sciences.*

Reising, D.V. (1993). *The effect of objective fault difficulty on human diagnostic reasoning with multiple faults.* Unpublished M.S.I.E. thesis. Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign, Urbana, IL

Sanderson, P.M. (1993). Designing for simplicity of inference in observational studies of process control: ESDA and MacSHAPA. *Proceedings of the Fourth European Meeting on Cognitive Science Approaches to Process Control (CSAPC' 93), Hillerod, Denmark, August 25-27.*

Sanderson, P.M. (1994). *Exploratory sequential data analysis: Software.* Engineering Psychology Research Laboratory Technical Report (EPRL-94-01). Urbana, IL: University of Illinois.

Sanderson, P.M. & Fisher, C. (1994). Exploratory sequential data analysis: Foundations. *Human-Computer Interaction*, 9.

Sanderson, P.M., Scott, J.J.P., Johnston, T., Mainzer, J., Watanabe, L., & James, J.M. (In press). *International Journal of Human-Computer Studies.*

Segal, L.D. (1994). Effects of checklist interface on non-verbal crew commication. NASA Contractor Report 177639. Moffett Field, CA.

Weitzman, E.M. & Miles, M.B. (1994). *Computer programs for qualitative data analysis.* Thousand Oaks, CA: Sage.

AVIATION TOPIC - SPEECH ACT TAXONOMY PC (ATSAT)

O. Veronika Prinzo
Federal Aviation Administration
Civil Aeromedical Institute

The Aviation Topics Speech Acts Taxonomy (ATSAT) is a tool for categorizing pilot/controller communications according to their purpose and for labeling communication errors. What makes the ATSAT different from other taxonomies is that FAA Air Traffic Control Order 7110.65 served to guide its development. Specifically, verbal communications that deviate from the standards specified in FAA Order 7110.65G (or suggested pilot communication in the Airman's Information Manual) can be identified and labeled, using the error codes provided in the ATSAT.

We have used the ATSAT to identify, classify, and code communication errors made by controllers and pilots during day to day field operations. We currently are investigating the effects that poorly constructed pilot messages transmitted during light and heavy traffic have on controller verbal communications and performance. The ATSAT will be used to identify, classify, and grade controller responses. By using the same procedures and tool to analyze communications, direct comparisons between controller phraseology usage in the field and during simulation can be made.

ATSATpc is a mouse-operated, Windows-based computer program. It is written in Visual Basic and requires SPSS for data analysis. ATSATpc consists of 5 main menus:

- File Information Menu
- Transmission Identification Menu
- Speech Act Category Menu
- Aviation Topic Menu
- Communications Error Menu

File Information Menu

The file information menu is used to select the transcription text file to open for data coding. Any ASCII text file with a .CMM extension can be analyzed. As shown in Figure 1, a 3-digit facility and sector code and a 2-digit controller code are typed into the appropriate box by the coder. After the enter button has been pressed, the program generates a window similar to Figure 2.

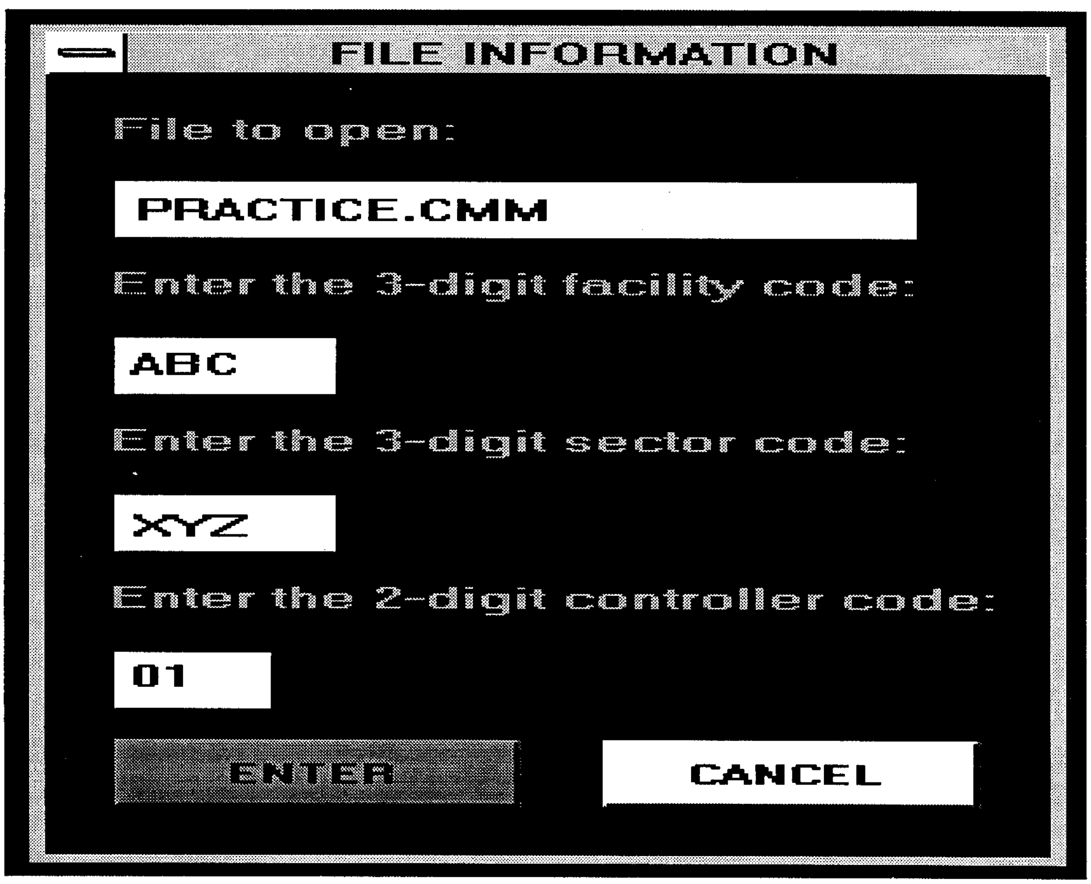


FIGURE 1: File Information Menu

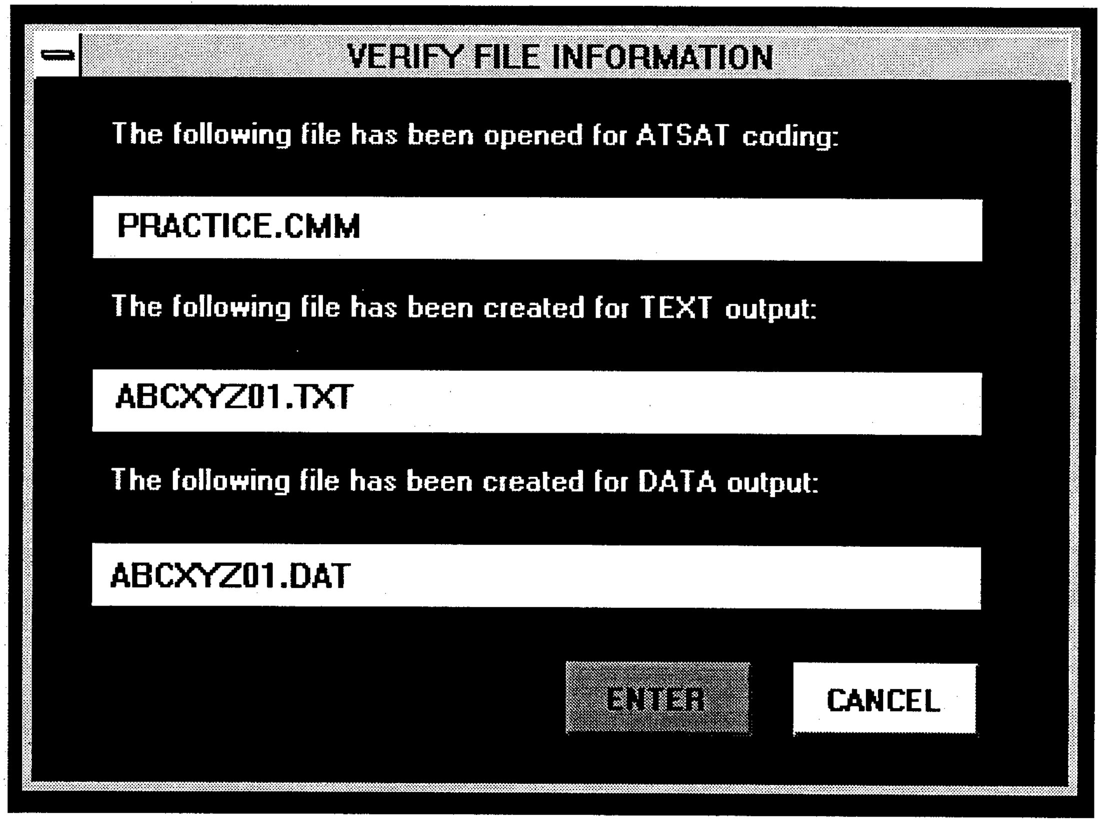


FIGURE 2: Window displaying input and output file names.

As shown in Figure 3, the ATSATpc creates 2 files: A text file that contains any key entries and general comments made by the coder, and a tab-delineated spreadsheet of the data set that can be exported to SPSS for statistical analysis. The original transcription text file is left unchanged by the program. Once the Enter button is pressed, the next menu is displayed with the contents of line 1 of the transcription file.

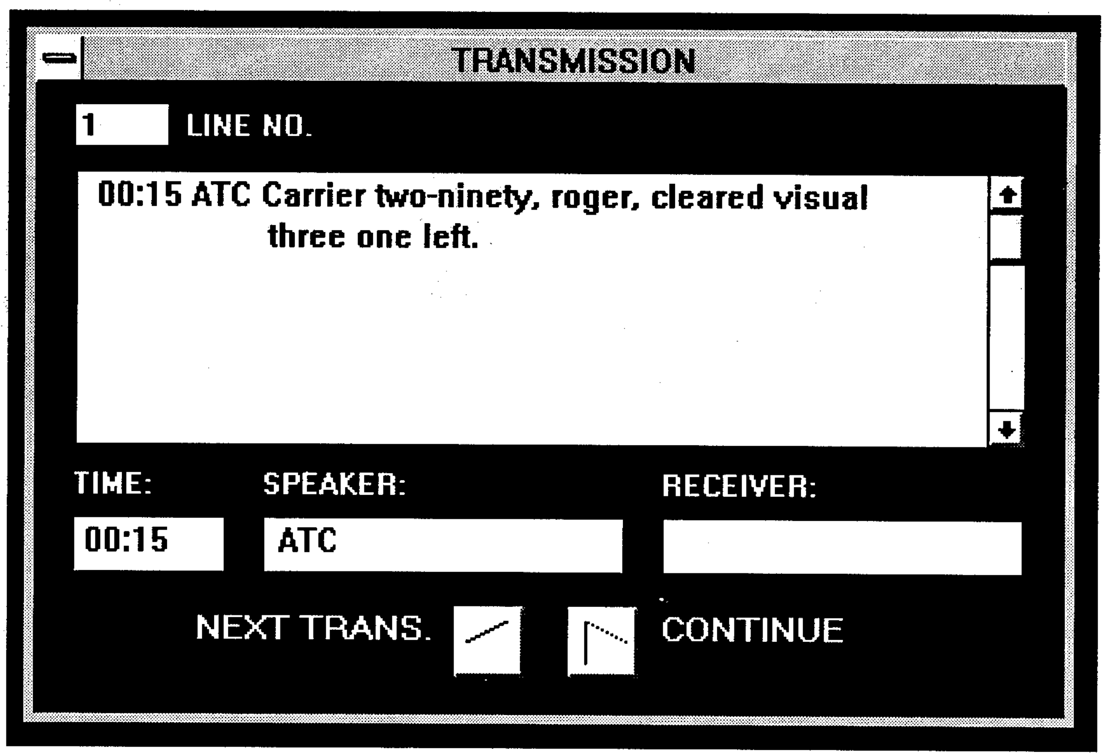


FIGURE 3: Transmission Identification Menu

Transmission Identification Menu.

At the Transmission Identification menu, each transmission is tagged with who generated the message and the intended recipient. The coder will identify the speaker and receiver of the transmission displayed in the dialogue box. The coder highlights the communication element in the transmission that corresponds with the word label "time," "speaker," or "receiver," places the cursor on that word label, and then double clicks the mouse button. The highlighted information is copied directly into the box beneath the word label and entered directly onto the spreadsheet. Pressing the continue button takes the coder to the Speech Act Category menu presented in Figure 4. Selecting the next transmission button causes the next transmission to appear in the dialogue box.

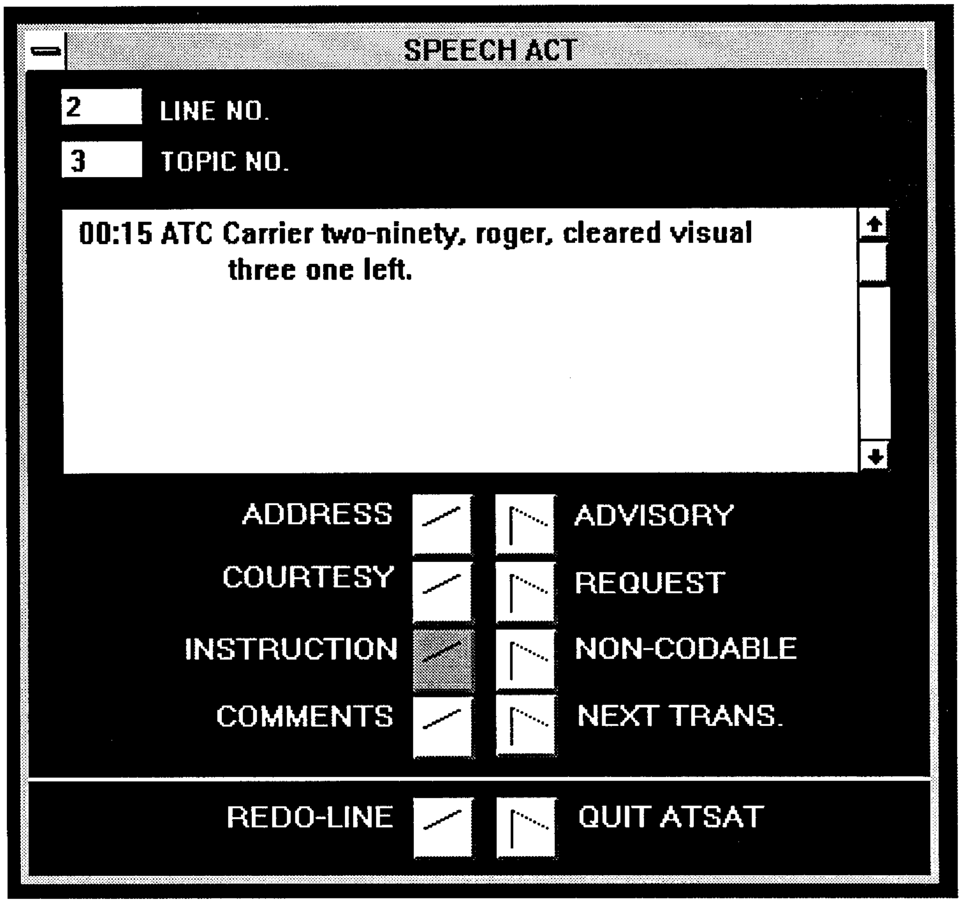


FIGURE 4: Speech Act Category Menu

Speech Act Category Menu.

The speech act category menu allows the coder to select and label a communication element by its purpose (what). A speech act is a single utterance which suggests an action. The speech act menu includes Address, Courtesy, Instruction, Advisory, Request, and Non-Codable. The Address is the who of the transmission. It references either an aircraft or the air traffic control facility position/sector. In addition to showing a level of respect, a Courtesy often signals the end of a dialogue between the air traffic controller and the pilot in much the same way that a good-bye signals the end of a telephone conversation. The Instruction, Advisory, and Request speech act categories represent what the communication element in the message is about - the action to be undertaken. They represent the "do something" "tell something" and "ask something" of an utterance. For example:

"Carrier two-ninety, roger, cleared visual three one left" contains three speech acts: Address - Carrier two ninety; Instruction - roger, Instruction - cleared visual three one left.

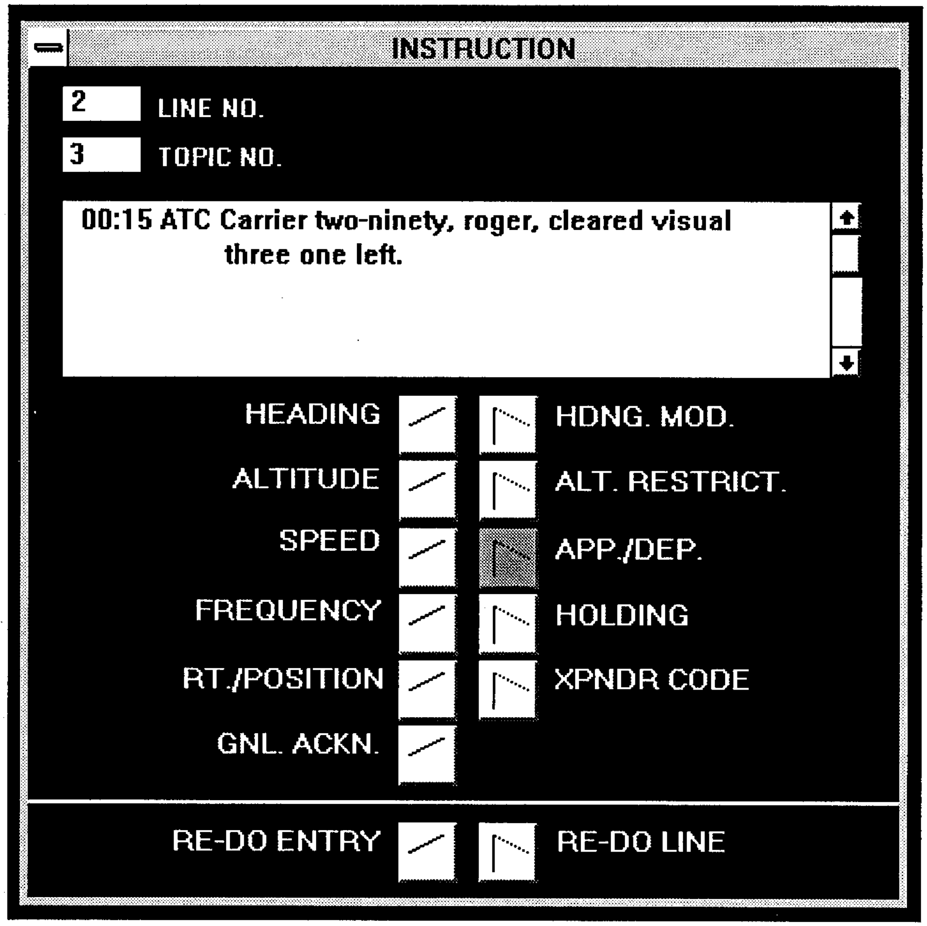


FIGURE 5: Aviation Topic Menu

Aviation Topic Menu.

The Aviation Topic places a constraint on the communication element by imposing a restriction on its identified speech act category (who, what). For example, there are only 2 types of aviation topics listed under the Address speech act category. There only can be 1 speaker and 1 receiver of a transmission. There are 3 types of aviation topics listed in the Courtesy speech act category: Thanks, Greetings, and Apology. The types of aviation topics listed in the Instruction, Advisory, and Request speech act categories are not exhaustive but represent the most frequently uttered messages that we heard from field tapes. The example of the earlier transmission has been embellished to include the types of aviation topics:

"Carrier two-ninety, roger, cleared visual three one left" contains three aviation topics: Address [Receiver] - Carrier two ninety; Instruction [Genl Ack] - roger, Instruction [App./Dep.] - cleared visual three one left.

As displayed in Figure 5, there are 11 different aviation topics that are listed for the Instruction speech act. In the example, the coder would have selected the speech act Instruction and then the menu displayed in Figure 5 would have appeared. The coder would select the aviation topic that represented the communication element and then decide if a communication error was present. If the coder indicates that an error is present then the menu displayed in Figure 6 would appear, otherwise the speech act menu appears.

Communications Error Menu.

The Communication Error menu is used by the coder to grade the contents of the communication element and label the detected message content errors and the delivery technique errors. The types of message content errors are grouped, sequential, omission, substitution, transposition, excessive verbiage, and partial readback. The example of the earlier transmission has been embellished to include the identified communication errors:

"Carrier two-ninety, roger, cleared visual three one left" contains one communication error: Address [Receiver] - Carrier two ninety; Instruction [Genl Ack] - roger, Instruction [App./Dep./O] - cleared visual three one left.

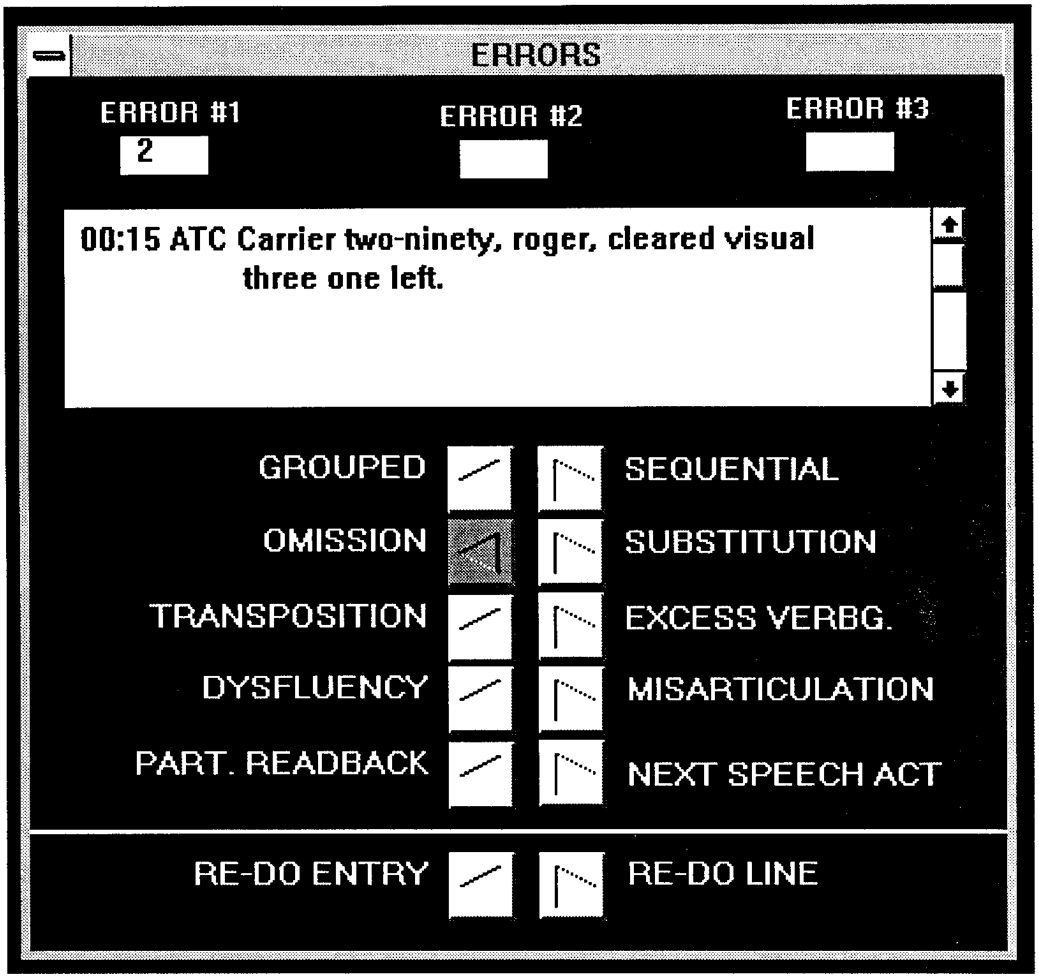


FIGURE 6: Communication Error Menu

Correct phraseology for the approach clearance is "... cleared visual *approach runway* three one left." Failure to include the words "approach" and "runway" as part of the clearance as required in the FAA Air Traffic Control Order 7110.65 results in nonstandard phraseology. The example is coded as an omission error.

The ATSATpc is a tool that uses the FAA Air Traffic Control Order 7110.65 to grade air traffic control and pilot communications. By using the same phraseology that controllers are required to use when speaking to pilots as the metric to grade their actual messages, the likelihood of comparing apples to oranges is eliminated. Subsequent analyses can determine where deficiencies occur and recommendations made to correct any carelessness on the part of the speaker. On the other hand, it may be that in spite of the speaker's best efforts to comply with FAA Air Traffic Control Order 7110.65, changes to the standard phraseology are warranted.

APPENDIX A

DISCOURSE PROCESSES

Editor's Note:
The manuscripts contained in this appendix are reprinted with written permission from the authors and publishers. We thank them for their help and support in making this report a more comprehensive repository of the methods and metrics used in voice communications.

International Journal of Aviation Psychology, Vol. 4, Helmreich, R.L., Anatomy of a system accident: The crash of Avianca Flight 052, pp. 265-284, (1994), with kind permission from Lawrence Erlbaum Associates, Inc.

Anatomy of a System Accident: The Crash of Avianca Flight 052

Robert L. Helmreich

The University of Texas at Austin

On January 25, 1990, Avianca Flight 052 crashed after running out of fuel following a missed approach to New York's John F. Kennedy Airport. Weather was poor on the East Coast of the United States that day, and the flight had experienced several holding patterns enroute from Medellín, Colombia, to New York. The accident is analyzed in terms of Helmreich and Foushee's (1993) model of crew performance and Reason's (1990) model of latent pathogens in system operations.

Although there is general consensus that flight crew behavior is implicated in more than two thirds of all air transport accidents and incidents (Helmreich & Foushee, 1993), it is also clear that pilot error is seldom the sole cause of an accident. This is borne out by the findings of the Canadian Commission of Inquiry into the Air Ontario crash at Dryden, Ontario, the most exhaustive investigation ever conducted into a single crash (Helmreich, 1992; Moshansky, 1992). What seemed to be a simple case of a tragically flawed pilot decision to take off with ice on the wings was shown after 3 years of investigation and more than 165 days of testimony to be a system accident to which regulatory, organizational, environmental, group, and individual factors contributed.

Although the U.S. National Transportation Safety Board (NTSB) does a commendable job of investigation, it seldom if ever has the resources to mount the kind of inquiry conducted by the Canadian Commission. The NTSB report on the Avianca Flight 052 (AV052) accident pinpointed a number of factors (including crew performance) that contributed to the crash (NTSB, 1991a). However, a number of additional pieces of evidence were uncovered in the course of litigation between the airline and the U.S. Gov-

Requests for reprints should be sent to Robert L. Helmreich, Department of Psychology, University of Texas, Austin, TX 78712.

ernment. Our article approaches the accident from a system and group perspective and utilizes several methodologies to attempt to explain the multiple causal factors at play on the night of January 25, 1990. The analysis was guided by Reason's (1990) notions of latent failures and resident pathogens in complex systems.

MODELS OF FLIGHT CREW PERFORMANCE

The model of crew performance proposed by Helmreich and Foushee (1993) was adapted from a more general model developed by McGrath (1964). The model identifies input factors that are present at the initiation of a flight, process factors that reflect the interpersonal and technical enactment of group tasks, and outcome factors that define multiple dimensions of success or failure on tasks undertaken. Critical to the model is the notion of feedback loops among the factors. Process factors influence not only outcomes but inputs to subsequent performance, and intermediate and final outcomes influence present and future processes and inputs. *Input factors* include national and organizational cultures and norms; organizational resources and practices, including training, support, and maintenance; environmental aspects, including weather, group structure, and composition; and individual characteristics, including personality, motivation, attitudes, and aptitude. Reason's (1990) concepts of latent failures and resident pathogens relate primarily to the input factors that define the operational shell within which group processes occur, although they can also influence group processes in a variety of ways.

METHOD

Three methods were employed to analyze the individual, group, and system aspects of this accident. The first method involved a review of documents and depositions generated during the discovery phase of litigation. These gave a picture of the organizational culture and practices, including the training of flight crews, dispatch practices, and maintenance. These data defined input factors that were potential influences on group processes of the flight crew.

The second method involved assessment of crew behaviors in terms of behavioral markers that were developed as part of the author's and his colleagues' research into the evaluation of crew performance. The Avianca crew was coded on the presence, absence, and valence of 52 specific behaviors; these data were compared with those of other accidents that have been analyzed using this approach.

The third method involved creating a data base of crew, air traffic control (ATC), and other aircraft communications. All communications from NTSB and Federal Aviation Administration (FAA) transcripts and the cockpit voice recorder (CVR) were broken into single utterances and put into the database. Each utterance was coded in terms of speech form (i.e., inquiry, observation, command, etc.) and classified in terms of content into Action Decision Sequences (ADS). The ADS is defined as all communications surrounding a particular course of action or situation (e.g., making an approach or evaluating fuel status). This analytic system described as microcoding was refined by Predmore (1991, 1993) and employed by him in the analysis of crew behavior in two United Airlines accidents and a number of experimental simulations.

SYNOPSIS OF THE ACCIDENT FLIGHT

AV052, a Boeing 707, crashed at 2134 EST on January 25, 1990, in a wooded residential area on Long Island while maneuvering for a second approach to New York's John F. Kennedy (JFK) airport. It was a scheduled flight from Medellín, Colombia. Of the 158 persons aboard, 73 were fatally injured.

Weather conditions were poor on the Eastern seaboard and the flight was placed in holding three times by ATC for a total of 1 hr, 17 min. While in the third holding pattern, the crew reported that they could not hold longer than 5 min and that they could not reach their scheduled alternate, Boston. On being cleared to JFK after this interchange, the crew executed a missed approach. While trying to return to the field, the airplane experienced a loss of power to all four engines as a result of fuel exhaustion and crashed approximately 16 miles from JFK.

RESULTS: ORGANIZATIONAL AND SYSTEM FACTORS

Avianca Management

Dispatch. The dispatching of the flight was deficient in a number of ways. The weather report provided to the crew was 9 hr old when the aircraft left Medellín. The dispatcher involved stated that aircraft were dispatched to New York without consideration of weather conditions. He also reported that Boston was always used as the alternate for New York, even if the weather was below minima. The company's own report on the accident described the state of dispatching in the organization:

> This chapter necessarily deals with the lack of real flight dispatchers in the company at the time of the accident. Only 3 dispatchers were truly qualified as such. The rest of the personnel was a group of persons better categorized as balancers lacking the background to function as dispatchers. This situation is

> the case throughout all the Avianca bases in the country. This condition should, of course, be supervised by the Office of the Director of Flight Operations, but in fact it was not done, and this situation was allowed to prevail for a long time in the condition previously described.
>
> *Medellín Flight Dispatcher for AV052.* This person was not qualified as a dispatcher and for that reason he could not adequately give the assistance needed by the crew of AV052, since he was unaware of a series of requirements that the flight should have met, which were not discussed with the crew. Likewise, the weather information was not brought up to date because neither this person nor the Bogotá dispatchers requested the new information that affected the flight and which would have provided the crew with a precise and more organized plan for the flight. (AV024792)[1]

With regard to the Operations Office supervision of dispatch, the report went on to state:

> This office did not furnish the up to date weather information needed to begin or plan the flight, either in Bogotá or Medellín. This factor was due to that staff's ignorance of the pertinent regulations at the time of the accident, because it did not have the required preparation or training to act as a duly qualified flight dispatcher. (AV024835)

Flight operations and flight training. A summary of the status of operations at Avianca is found in the company's investigation of the accident:

> About 1960, the company introduced the B-707 and 320C[,] subsequently the B-727, the B-737, the B-747, and finally today the B-767. For this aircraft and this type of operation, so far as its flight operations department was concerned, the company *retained* the same operations manual from the conventional aircraft period with some small modifications and additions, thus remaining years behind in the updating of the same manual, which *is not consistent* at the present time with the airline operations that the company carries out today. In other words, the company does not have an airline policy for its operations that is defined by the company itself, and this permits improvisation in operations with the consequent decline in air safety. Proof of this is how only after the accident involving AV052 in New York on January 25, 1990, some policies to be followed with regard to flight operations are just being worked out. (AV024786-787)

Flight manuals available to B-707 crews were obsolete and did not include Boeing safety bulletins regarding minimum fuel. At this time, Avianca did not provide crewmembers with initial or recurrent training in Crew Resource

[1]References labeled "AV" refer to identification numbers of documentary evidence for litigation in U.S. federal courts.

Management (CRM). Pilots received only a single simulator period during the course of a year.

Minutes of the airline's Committee on Air Safety show recognition of deficiencies in training through the following comments:

> The Chief of the Committee emphasized the need to improve land-based training for crew members, and two proposals were put forward: (a) Close escope [Avianca's flight training establishment] and provide land-based training through accredited training schools; (b) Improve escope by means of additional capital and a reorganization, since its newest training tools are 17 years old and the staff resources are extremely limited. (AV023387)

The internal report of the accident investigation further pinpoints:

> . . . lack of a definite policy on Avianca air operations on the part of the Office of the Vice President of Operations and the Office of the Director of Operations, on which the crew could have relied to get an evaluation of its operation en route to New York. This would have made it possible to have a route profile for the flight in question, with the company's specific recommendations for its completion by the crew under various circumstances and events, which would have served as their guide for conducting the flight with the various operating alternatives most suitable to the company and the crew. (AV024822)

Two bulletins issued in 1985 indicated a continuing problem with adherence to safety related procedures at Avianca. The first, from the Director of Flight Training, addressed the fact that checklists were not being completed properly:

> It has come to my attention through several sources, one of them by a special written report from the Boeing instructors who recently were here, in which they complained, and other reports agreed, that checklists were not being read. As you well know this omission shows carelessness on the part of the Captain or Co-pilot and Engineer since not reading the checklist or doing it from memory is the most serious infraction that crewmembers, or member, can commit.
>
> For this reason let me put this in the form of an order, the reading (not memory) of the Check Lists. This office will use every means of control for the accurate fulfillment of this strict order, with disciplinary actions for those who fail to comply. (AV018667)

The CVR transcript indicates that the crew did not complete the B-707 Normal Checklist for Landing correctly. Thus, the informal, operational culture apparently did not reflect the organization's stated concerns.

The second bulletin was issued by the Colombian Department of Civil Aeronautics. It was distributed to Avianca pilots and stated:

This circular has as its object to CARRY OUT GUARANTEES OF FLIGHT SAFETY, reminding the proper employees of radiotelephone procedures which are being done incorrectly by some crewmembers. THE USE OF NON STANDARD PHRASEOLOGY can cause misunderstanding; incidents and accidents have happened in which this has been a contributing factor, confusion caused by POOR PHRASEOLOGY. (AV018624)

Information regarding the state of training and operations at Avianca is found in the transcript and recordings of a conference on human factors and CRM conducted for Avianca personnel. The conference provided a review of critical issues in CRM and discussed them in the context of accidents at Avianca:

Finally, in our company the last four jet plane accidents (Barranquilla, Cucuta, Madrid, and New York), had to do with airplanes in perfect flight condition, aircrew without physical limitations and considered of average or above average flight ability and still the accidents happened.... which leads us to believe that the possible causes were: lack of decision making ability (or inadequate ability in this regard), the lack of coordination in the cockpit, the lack of command, leadership, communication, or teamwork. This suggests that traditional training is not focused toward these areas. The errors involved in the majority of accidents are caused by the failure of all crewmembers to make use of all available resources. Therefore training must cover these new needs to teach crewmembers the correct way to operate as a flight team. (AV020886)

Later in the conference, the discussion focused on communications skills as an essential means of maintaining situational awareness; this was identified as a problem in three Avianca accidents, including AV052:

Communications Skill. A flightcrew spends much of the time communicating. This is the most essential factor for good performance in the cockpit. If communication among crewmembers is effective, performance in the cockpit will be improved and the crew can reach and maintain a high level of situational awareness. If the communication is not effective, mistakes and erroneous interpretations will occur and situational awareness will be lost. The consequences can be serious and frequently disastrous, for example, 747 in Madrid, 1716 at Cucuta, 2016 at JFK. (AV010918)

Later discussion returned to the Avianca B-747 crash in Madrid, which involved warnings from the Ground Proximity Warning System (GPWS) and ineffective communication by the first officer to make the captain aware of the dangerous situation they were in:

Madrid Avianca. The co-pilot [sic] was right, but they died because the captain kept on believing in his false situational awareness. When the co-pilot asked questions, his implied suggestions were very weak. The captain's reply was to ignore him totally. Perhaps the co-pilot did not want to appear rebellious, questioning the judgment of the captain, or he did not want to play the fool because he knew that the pilot had a great deal of experience flying in that area. The co-pilot should have advocated for his own opinions in a stronger way, giving clues to the pilot so he could realize that his situational awareness was low. (AV020921)

Maintenance. The captain accepted (or was subtly pressured by organizational norms to accept) an aircraft that had several maintenance deficiencies. The autopilot was not working and had a number of maintenance write-ups in the preceding month. On the day of the accident, this necessitated hand-flying, with an associated increase in workload, fatigue, and stress. The day before the crash, the malfunctioning autopilot had been described in a logbook write-up as "abnormal and dangerous" by the second officer of AV052. Another write-up that month had asked for investigation of the "implications of a flight of more than two hours with an autopilot that is inoperative."

Maintenance standards are captured in the company's report on the accident:

A. According to analysis in the NTSB laboratory, the flight data recorder was found inoperative due to corrosion and incorrect installation of the magazine or recording tape. This situation prevented the investigation from having the element of proof regarding the flight parameters of the HK2016. This condition reflected the lack of strict adherence to the maintenance schedule of this type of equipment.

B. The presence of a decaying cardboard box inside one of the fuel tanks also reflected a situation of neglect on the part of those responsible for these (fuel) systems and for their cleanliness, so necessary for any airline in its operations. (AV024812)

It cannot be ascertained whether the organization routinely forced captains to accept deficient aircraft or whether the captain simply failed to be concerned with the maintenance problems of the airplane.

ATC

It is evident that ATC did not realize the severity of AV052's fuel state and hence did not treat the flight as being in an emergency. Three air traffic controllers who handled the flight testified that they did not perceive a crisis aboard AV052 when the first officer made reference to the aircraft's fuel state.

It is clear from review of ATC transcripts of communications on frequencies assigned the flight that there was a great deal of information about lengthy holds, possible wind shear, poor visibility, and flights diverting to alternates on the party line.[2]

ATC assumed from the crew's communication between 2044 and 2047 that, if cleared from the holding pattern within 5 min, AV052 could proceed with routine handling. The flight received the service it requested. The controller also had the expectation that if the service delivered was not as expected and desired, the crew would speak up and make its status, preferences, and intentions known to subsequent controllers—which it did not do. The crew's disregard for FAA, Avianca, and International Civil Air Organization procedures, shown in their failure to declare an emergency and to make its intentions known, diminished any sense of urgency on the part of ATC.

The crew could have recognized that they were being given routine radar service to place them in sequence with other aircraft on approach to JFK from the ATC party line, from the vectors they were given, and from their own distance measuring equipment. It was also apparent from this information that they were not being given direct routing to JFK and were not being placed ahead of other aircraft. The priority requested and granted was in departing the CAMRN holding pattern. The crew could have obtained direct routing into JFK by rejecting the clearance that was delivered.

Additional information regarding the routine treatment of the flight was available from the vectors issued by ATC, which took the aircraft away from the approach. Again, the crew could have rejected this clearance and declared the need to land immediately under emergency conditions.

In addition to the failure to use required, standard terminology to communicate flight status, the information regarding fuel state and the need for "priority" was communicated in an offhand manner. This, combined with the first officer's excellent, unaccented English and the monotone voice with which the information was transmitted, misled controllers.

Table 1 shows communications from AV052 to ATC that deal specifically with the status of the flight. Examination of these utterances, which were issued to four different controllers, gives an indication of why ATC perceived transactions with the flight to be of a routine nature. There was an inquiry about the status of Boston and a follow-up to this query at 2005:37. There were no further communications regarding status until four transmissions between 2044 and 2047. There were no further transmissions regarding aircraft status from AV052 until 2124, a period of 38 min during which no effort was made to update ATC regarding the situation. None of these communications made clear the fuel situation of the flight, stated intentions, or offered alternatives to ATC. A number of the communications are phrased in a tentative manner rather than as statements of necessity or urgency. For example, the communication at 2044:50 states "... WELL I THINK WE

[2] *Party line* refers to the information to be obtained by monitoring communications between ATC and other aircraft on the same radio frequency.

TABLE 1
Avianca Flight 052 Communications to ATC Regarding Flight Status

Time	To	Utterance
Communications to ATC Regarding Status Prior to Missed Approach		
2003:26	R59	YOU HAVE ANY INFORMATION ABOUT DELAYS AT BOSTON
2005:37	R59	DID YOU ASK ABOUT ANY DELAY UH, AT BOSTON OR ARE WE GOING TO APPROACH TO KENNEDY
2044:50	R67	ZERO TWO ZERO FIVE AHHHH WELL I THINK WE NEED PRIORITY WE'RE PASSING (unintelligible).
2046:03	R67	YES SIR AH WE'LL BE ABLE TO HOLD ABOUT FIVE MINUTES THAT'S ALL WE CAN DO.
2046:13	R67	OH WE SAID BOSTON BUT AH IT IS AH FULL OF TRAFFIC I THINK.
2046:24	R67	IT WAS BOSTON BUT WE WE CAN'T DO IT NOW WE, WE, DON'T, WE RUN OUT OF FUEL NOW.
2046:24–2124:07: No communications regarding status		
Communications to ATC Regarding Status After Missed Approach		
2124:07	TWR	THAT'S RIGHT TO ONE EIGHT ZERO ON THE HEADING AND AH WE'LL TRY ONCE AGAIN, WE'RE RUNNING OUT OF FUEL.
2125:07	FV	CLIMB AND MAINTAIN THREE THOUSAND AND UH WE'RE RUNNING OUT OF FUEL SIR.
2126:41	FV	I GUESS SO THANK YOU VERY MUCH.
2130:40	FV	AH NEGATIVE SIR WE WE'RE JUST RUNNING OUT OF FUEL WE OKAY THREE THOUSAND NOW WE COULD.
2132:51	FV	AVIANCA ZERO FIVE TWO WE JUST AH LOST TWO ENGINES AND WE NEED PRIORITY PLEASE.

Note. R59, R67 = radar controllers; TWR = JFK control tower; FV = final vector controller.

NEED PRIORITY . . . " Later, with regard to their alternate, AV052 states that ". . . AH IT IS AH FULL OF TRAFFIC I THINK . . . " It can also be seen that statements regarding the situation are appended to other communications in several cases rather than appearing as single messages. For example, at 2124:07, ". . . AH WE'LL TRY ONCE AGAIN, WE'RE RUNNING OUT OF FUEL . . . ," and at 2125:07, ". . . CLIMB AND MAINTAIN THREE THOUSAND AND UH WE'RE RUNNING OUT OF FUEL SIR . . . ," and so forth.

Group Processes and Crew Behavior

The human factors of flight crew performance have come to be classified under the label of CRM (Helmreich & Foushee, 1993). *Appropriate CRM* is defined as the utilization of all available resources, which includes other crewmembers; manuals and other documentation; dispatch, flight service stations and flight-following services; and ATC and the ATC party line. Crew behaviors associated with effective flightdeck management include (a) leadership and task orientation, (b) team formation and maintenance, (c) acquisition and exchange of appropriate information, (d) problem solving, (e) decision making, and (f) maintaining situational awareness.

The AV052 Crew

As an example of an active failure, the crew disregarded Avianca and FAA procedures by failing to give a minimum fuel advisory. Even though their situation was becoming increasingly serious, the crew failed to follow standard procedures, including the declaration of an emergency, to ensure a prompt landing with adequate reserve fuel.

As mentioned in the preceding section, communications on frequencies monitored by AV052 showed that there was a great deal of information available on flights holding and diversions because of fuel state. Yet during the time period covered by the CVR, there was no discussion among the crew regarding alternative courses of action, such as selecting a new alternate and diverting. In addition, there was no discussion of actions to be taken in the event of encountering reported wind shear or of what should be done in the case of a missed approach at JFK.

External resources such as dispatch services or the Miami Flight Service Station were not employed to obtain current information on weather, delays, or available alternates. At 2044 EST, when mention was first made of a need for "priority," the situation was already critical. The crew was certainly aware that in the event that JFK should close or a missed approach should be necessary, their fuel state would be critical and could result in a crash.

During the period of time covered by the CVR, there were few intracockpit communications regarding the worsening fuel state. None of the sparse communications associated with fuel state included a discussion of actual status or addressed contingency planning.

Behavioral markers of crew performance. Transcripts of 10 aircraft accidents were reviewed by expert raters to determine whether the behavioral markers that are used to evaluate crew performance in line operations and simulations could be evaluated from CVR records taken under emergency conditions.[3] Each of the markers represents a behavior found to be positively associated with effective crew performance (e.g., Helmreich & Foushee, 1993; Helmreich, Wilhelm, Kello, Taggart, & Butler,1991). These accidents included AV052 as well as several other crashes (e.g., United Airlines Flights 232 and 811; NTSB, 1990b; 1990c) where the crew's perfor-

[3]The behavioral markers are presented in detail in FAA Advisory Circular 120-51a, *Crew Resource Management* (Washington, DC: Author).

mance was deemed to be exemplary by the NTSB (Jones, 1993). A coding system was employed in which each of the 52 markers was assigned a score of 1 (*present*), –1 (*absent or ineffective*), or 0 (*inapplicable or indeterminate*). These ratings were summed, with a possible range of 52 to –52. Results of the summed coding are shown in Figure 1. Comparison of sum scores for accidents where crew performance was seen as effective and those where human factors deficiencies were noted yields a highly significant t test, $t(7) = 7.9$, $p < .001$. The AV052 crew had a score of –38, indicating very ineffective use of human factors concepts.

The microcoding of crew communications from the CVR defined six action decision sequences (ADSs) that were or should have been present in the verbal exchanges in the cockpit. Table 2 lists those specified for AV052. Figure 2 shows the number of communications associated with each of the ADSs defined for the flight.

What is most significant about this breakdown of communications is the low percentage addressed to the critical problem of fuel state. Only 19 utterances were made on this topic during the period of more than 30 min covered by the CVR. Equally notable is the complete absence of communications to the cabin regarding preparations for a possible emergency landing or crash.

Overall, the total amount of communication within the cockpit was very low. In his analysis of effective crews in extreme emergencies, Predmore (1991; 1993) found high levels of information exchange (ranging from about 20 utterances per min during routine operations to averages over 35 per min, with a peak of up to 60 per min during the United 232 accident at Sioux City, IA; NTSB, 1991a). This crew was not exchanging critical information on flight status and on possible courses of action. Figure 3 shows the distribution of communication across time.

It can be seen that little attention was paid to the worsening fuel state and much of that was following the missed approach when the flight's situation was *in extremis*.

Captain. The captain's behavior reflected reactive rather than proactive leadership. External resources such as updated weather from Bogotá, dispatch services, and various flightwatch services were not utilized, al-

TABLE 2
Definition of Action Decision Sequences in Avianca Flight 052

Initial approach to JFK
Fuel state/contingencies
Missed approach
Cabin preparation
Weather/holding
Nonoperational communications

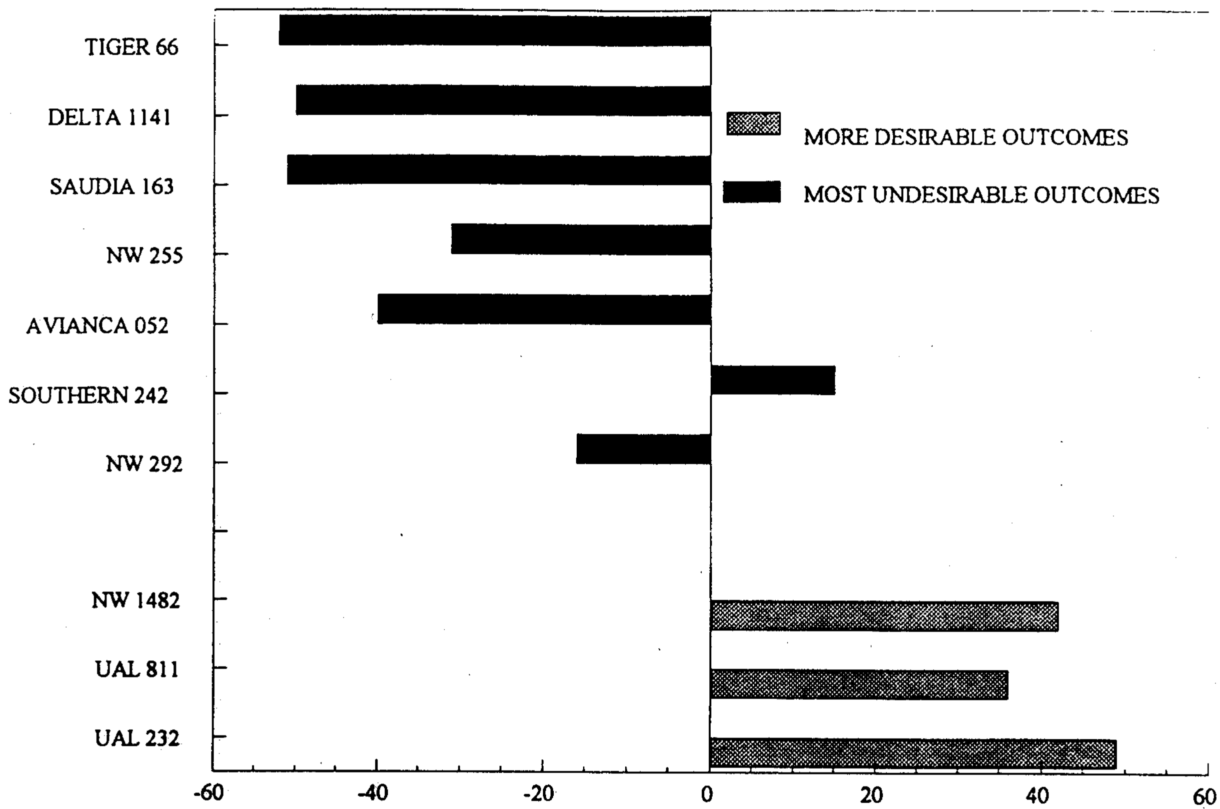


FIGURE 1 Summed scores based on the presence or absence of 52 behavioral markers (from Jones, 1993).

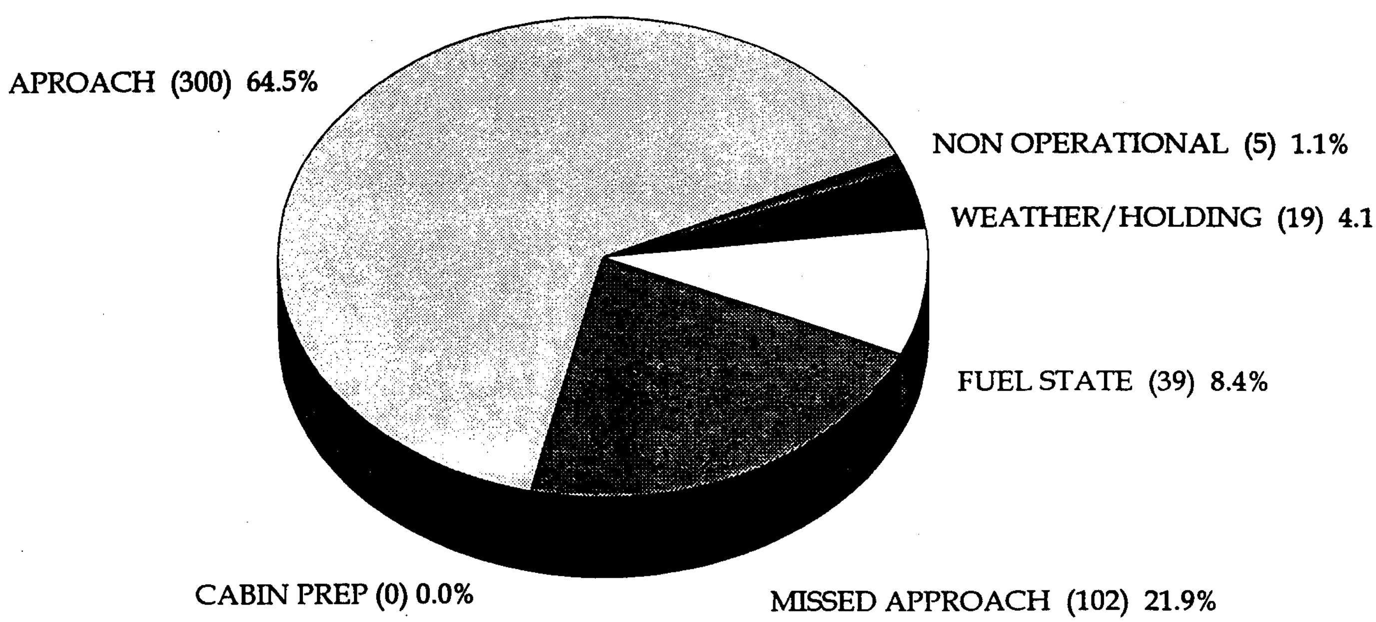


FIGURE 2 Distribution of action decision sequences in Avianca Flight 052.

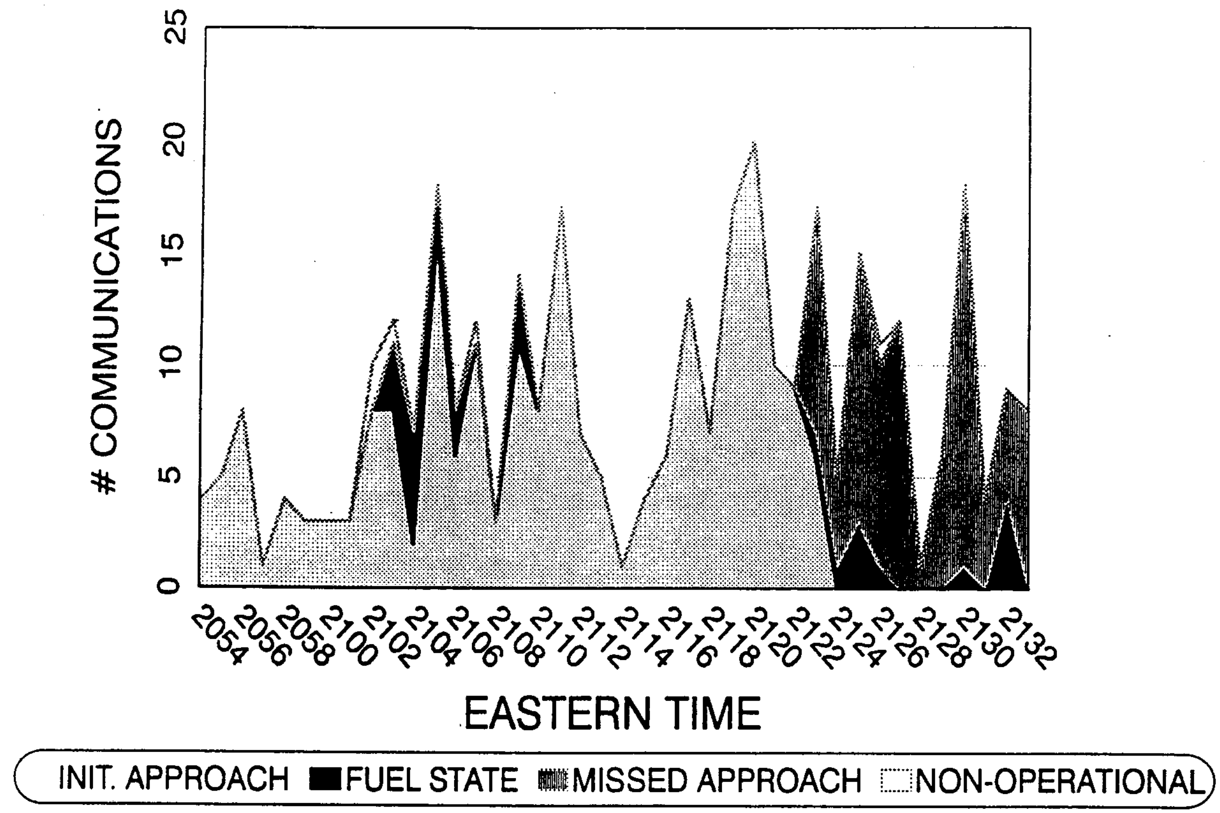


FIGURE 3 Distribution of action decision sequences over time from CVR.

though seeking current information was routinely expected of crews flying to the United States from Colombia. The captain also failed to utilize internal resources in the form of the other crewmembers to help in situation assessment and planning. He failed to discuss contingencies in the event of wind shear or a missed approach at JFK and did not communicate his overall intentions to the other crewmembers. Failure to alert the cabin to the possibility of a crash landing is a further indication of a lack of leadership.

Examination of cockpit communications provides data on the captain's management of the cockpit. Inquiry (or seeking information from other crewmembers) regarding current status is a means of achieving and maintaining situational awareness. Table 3 lists the captain's instances of inquiry and shows that his queries were directed almost entirely toward finding out what ATC was saying and the current configuration of the aircraft.

TABLE 3
Inquiry by the Captain Regarding Current Status

Time	*Utterance*
2054:49	TWO TWENTY?
2055:07	HOW MUCH?
2055:08	TWO TWENTY?
2056:13	TWO TWENTY, CORRECT?
2056:28	WHAT IS HE SAYING WIND SHEAR?
2102:59	WHAT HEADING DID YOU SAY TO ME ZERO FORTY?
2104:59	WHAT HEADING DO YOU HAVE OVER THERE?
2105:11	WE PASSED ALREADY NO?
2105:34	TWO WHAT?
2105:39	WHAT HEADING HE PROVIDE US?
2105:52	HEY UNDERSTAND THAT NOSE MUST BE MAINTAINED AS LOW AS POSSIBLE
2107:56	WELL DO YOU WANT SET IT SYMMETRICALLY?
2111:49	DID YOU ALREADY SELECT FLAPS FOURTEEN NO?
2112:52	HOW MANY MILES IS THAT THING LOCATED?
2120:21	ARE WE CLEARED TO LAND NO?
2123:20	WHERE IS THE RUNWAY?
2123:23	THE RUNWAY WHERE IS IT?
2124:17	WHAT DID HE SAY?
2124:26	DID YOU TELL HIM?
2125:20	WHAT ZERO EIGHTY?
2125:28	DID YOU ALREADY ADVISE THAT WE DON'T HAVE FUEL?
2126:46	WHAT DID HE SAY?
2130:25	WHAT HEADING TELL ME?
2130:50	TELL ME–
2130:53	ARE THE FLAPS AT FOURTEEN?
2130:56	TELL ME HEADING WHAT?
2131:22	THREE SIXTY NO?
2133:22	DID YOU SELECT THE ILS?[a]

[a]ILS = instrument landing system.

The captain made no efforts to clarify the overall situation or to determine what actions were needed to accomplish a safe landing. These communications further indicate that (a) his command of English was poor, given the repeated need to clarify ATC communications; and (b) the first officer was not keeping the captain fully abreast of ATC communications directed at AV052 and heard on the party line. As the situation deteriorated, he became less aware of communications surrounding him as shown in Table 1 and commented at 2117:55, "TELL ME THINGS LOUDER—I'M NOT HEARING THEM."

There were also problems with the captain's technical performance and procedures in executing the approach to JFK. He failed to return to the glideslope after descending below it and being repeatedly advised of his deviation by the first officer. Flying on the glideslope should have made a landing possible, because the runway would have been visible at decision height. The captain also failed to react manually or verbally to 15 repeated GPWS alerts. Finally, although the fuel state was *in extremis* at the moment of the missed approach at JFK, the captain failed to initiate an immediate return to the runway for landing.

First officer. The first officer was inexperienced overall and particularly in the B-707, with fewer than 60 hr in this aircraft type. He disregarded the captain's order to declare a fuel emergency. As the aircraft was approaching JFK, he did not accurately communicate to the captain the wind shear information given by ATC, and incorrectly reported that the flight was being given priority. After the go-around, he reported falsely to the captain that he had declared an emergency and later when the captain ordered him to tell ATC that ". . . WE DON'T HAVE FUEL," instead stated at 2125 that ". . . WE ARE RUNNING OUT OF FUEL SIR." At 2125:15, he acknowledged a climb to 3,000 ft and a heading change to 080°. He reported the altitude change correctly to the captain but indicated that the heading should be "HUNDRED AND EIGHTY" degrees—a course away from the airport. The captain had apparently heard the heading correctly, because he inquired, "WHAT ZERO EIGHTY" at 2125:20, but received again the 180° heading from the first officer. At 2125:29, the first officer incorrectly gave the captain the 180° heading for the third time.

Another example of a failure in communication is found in an exchange with ATC and the captain following the missed approach. ATC asked about vectors for return to JFK as follows: ". . . I'M GUNNA BRING YOU ABOUT FIFTEEN MILES NORTHEAST AND THEN TURN YOU BACK ONTO THE APPROACH. IS THAT FINE WITH YOU AND YOUR FUEL?" The first officer replied, without consulting the captain, "I GUESS SO THANK YOU VERY MUCH." When the captain asked what ATC said, the first officer's reply failed to relay the query about fuel state, instead commenting, "THE GUY IS ANGRY."

Second officer. The second officer failed to provide the captain with continuing information on the worsening fuel state. The surviving flight attendant testified that the flight engineer mentioned three alternates—Boston, Philadelphia, and Dulles—while the attendant was in the cockpit. Later, on the first approach to JFK, he read from the manual's instructions for a missed approach with minimal fuel aboard. He did not, however, verbally communicate the urgency of the situation and the need to land on the first approach. Following the missed approach, according to the flight attendant's deposition, instead of speaking up regarding the gravity of the situation, he indicated the crisis to the flight attendant by pointing to the empty fuel gauges and making a gesture representing the cutting of a throat to indicate that the plane was about to crash.

IMPLICATIONS AND APPLICATIONS

There is no question that the crew's behavior showed failures in CRM, adherence to procedures, and technical performance. However, it would be wrong to stop with the conclusion that pilot error caused the accident. The latent failures in the organization—including training, operational procedures and manuals, crew pairing, dispatch, and maintenance—created a window of opportunity for an accident to occur. For example, the crew's failure to show any reaction to 15 occurrences of the GPWS suggests not only a failure in vigilance but also an organizational failure to provide adequate training regarding the use of this critical warning system.

Cultural Factors

Behaviors that may seem inexplicable to U.S. aviators, such as the failure to advocate alternative courses of action to the captain or to question ATC instructions, could reflect characteristics of the crew's national culture. Hofstede (1980) isolated four dimensions of culture in a study of work values in 52 countries. Three of these seem relevant to this accident: power distance (PD), individualism–collectivism, and uncertainty avoidance (UAV). Cultures that are high in PD perceive large distances between subordinates and leaders and show a reluctance to question the actions and decisions of superiors. Collectivist cultures value in-group harmony and show a reluctance to take actions that might disrupt group relations. The third dimension, UAV, reflects an unwillingness to change and a need to avoid uncertainty. Colombia scored high in UAV, whereas the United States was below the median. Merritt (1993) and Merritt and Helmreich (in press) have found that the first two of these three dimensions are reflected in crew members' attitudes regarding flightdeck management using data collected with the Cockpit Management Attitudes Questionnaire (Helmreich, Wilhelm, & Gregorich, 1991). Colombia, in contrast with the United States, scored strongly collectivist and high in PD.

The crew's behavior appears more understandable when viewed in terms of Hofstede's cultural dimensions. The high PD of Colombians could have created frustration on the part of the first officer because the captain failed to show the kind of clear (if not autocratic) decision making expected in high-PD cultures. The first and second officers may have been waiting for the captain to make decisions, but still may have been unwilling to pose alternatives. From the framework of an individualistic culture, the controllers would have expected the Colombians to declare an emergency and escalate their demands if their needs were not met. There are three reasons why they may not have done this. First, there is a natural modesty in collectivist cultures and an unwillingness to place themselves in front of others. The crew may have believed that other crews were equally in need of immediate attention and may have been unwilling to "jump the queue" by declaring an emergency. Second, the first officer may have felt uncomfortable with the prospect of confronting the captain with the seriousness of the situation and, hence, may have abbreviated and de-emphasized ATC communications. Third, in coming from a culture in which group harmony is valued above individual needs, there was probably a tendency for the crew to remain silent while hoping that the captain would "save the day." Instances have been reported in other collectivist, high-PD cultures where crews have chosen to die in a crash rather than disrupt group harmony and authority and bring accompanying shame upon their family and in-group.

High UAV may have played a role by locking the crew into one course of action and preventing discussion of alternatives and review of the implications of the current course of action. The crew may have preferred to maintain Boston as their alternate to the ambiguity of choosing another. High UAV is associated with a tendency to be inflexible once a decision has been made as a means of avoiding the discomfort associated with uncertainty.

Had ATC been aware of cultural norms that can influence crews from other cultures, they might have communicated more options and queried the crew more fully regarding the flight's status. Although there is no regulatory requirement for such actions by ATC, training in cultural factors for controllers who deal with a large number of foreign flights could enhance the safety of the system.

Although the hypothesis cannot be proved, the possibility that behavior on that night was dictated in part by norms of national culture cannot be dismissed. It seems likely that national culture may have contributed to inflexible decision making, that weak leadership may have been exacerbated by a normative reluctance to question that leadership, and that the need to maintain group harmony may have inhibited crewmembers from presenting their concerns and suggestions. Finally, mistaken cultural assumptions arising from the interaction of two vastly different national cultures may have prevented effective use of the air traffic system. Johnston (1993), Merritt (1993) and Merritt and Helmreich (in press) have suggested that North American approaches to CRM training may not be applicable in many cultures. This raises the important research question of how to measure significant cultural differences and how to adapt training to reflect them (Merritt & Helmreich, in press).

When the array of latent failures surrounding this accident is recognized, it becomes clear that this is a system accident. If we are to learn from the mistakes of the past and prevent similar human-factors accidents in the future, it is necessary to look beyond the proximal behavior of crews in emergencies. Understanding the input factors (including such broad dimensions as national and organizational culture) that shape group process is essential. Only with such knowledge can proactive steps be taken to enhance system safety.

The use of systems approaches to CRM training is becoming widespread. Analyses of accidents and incidents using the approach employed herein may enhance human-factors training by sensitizing crew members to the broader context in which they operate.

ACKNOWLEDGMENTS

Supporting resesearch for this article was sponsored by NASA Ames Research Center Cooperative Agreement NCC2-286 and Federal Aviation Adminstration Grant 92-G-017 to Dr. Robert Helmreich.

Documentary evidence regarding this accident was collected while the author was a consultant to the U.S. Department of Justice (DOJ).

Dr. Steven Predmore, now of the National Transportation Safety Board, assisted in the original microcoding of the cockpit voice recorder. Thanks are also due to Robert Gross and James Wilson of the Torts Branch of the Civil Division of the DOJ.

REFERENCES

Helmreich, R. L. (1992). Human factors aspects of the Air Ontario crash at Dryden, Ontario: Analysis and recommendations. In V. P. Moshansky (Commissioner), *Commission of inquiry into the Air Ontario accident at Dryden, Ontario: Final report. Technical appendices.* (pp. 319–348). Toronto, Ontario: Minister of Supply and Services.

Helmreich, R. L., & Foushee, H. C. (1993). Why crew resource management? The history and status of human factors training programs in aviation. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.). *Cockpit resource management.* (pp. 3–45). San Diego: Academic.

Helmreich, R. L., Wilhelm, J. A., & Gregorich, S. E. (1991). *Revised version of the Cockpit Management Attitudes Questionnaire (CMAQ) and CRM seminar evaluation form.* (Available from Robert L. Helmreich, Department of Psychology, University of Texas, Austin, TX 78712)

Helmreich, R. L., Wilhelm, J. A., Kello, J. E., Taggart, W. R., & Butler, R. E. (1991). *Reinforcing and evaluating Crew Resource Management: Evaluator/LOS instructor manual.* (Available from Robert L. Helmreich, Department of Psychology, University of Texas, Austin, TX 78712)

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values.* Beverly Hills: Sage.

Johnston, N. (1993). CRM: Cross-cultural perspectives. Keeping CRM is keeping the flight safe. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 367–397). San Diego: Academic.

Jones, S. (1993). Human factors information in incident reports. In R. S. Jensen (Ed.), *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 567–572). Columbus: Ohio State University.

McGrath, J. E. (1964). *Social psychology: A brief introduction.* New York: Holt, Rinehart & Winston.

Merritt, A. (1993). Cross-cultural attitudes of flight crews. In R. S. Jensen (Ed.), *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 561–566). Columbus: Ohio State University.

Merritt, A. C., & Helmreich, R. L. (in press). Human factors on the flightdeck: The influence of national culture. *Journal of Cross-Cultural Psychology.*

Moshansky, V. P. (1992). *Commission of inquiry into the Air Ontario accident at Dryden, Ontario: Final report* (Vols. 1–4). Toronto, Ontario: Minister of Supply and Services, Canada.

National Transportation Safety Board. (1991a). *Aircraft accident report: Avianca, Boeing 707-321B, HK2016 fuel exhaustion Cove Neck, New York, January 25, 1990* (Rep. No. NTSB-AAR-91-04). Washington, DC: Author.

National Transportation Safety Board (1991b). *Aircraft accident report: United Airlines Flight 232, McDonnell Douglas DC10-10, Sioux Gateway Airport, Sioux City, Iowa, July 19, 1989* (Rep. No. NTSB-AAR-90-06). Washington, DC: Author.

National Transportation Safety Board (1991c). *Aircraft accident report: United Airlines Flight 811, Boeing 747-122, N4713U, Honolulu, Hawaii, February 24, 1990* (Report No. NTSB-AAR-90-01). Washington, DC: Author.

Predmore, S. (1991). Microcoding of cockpit communications in accident analyses: Crew coordination in the United Airlines Flight 232 accident. In R. S. Jensen (Ed.), *Proceedings of the Sixth International Symposium on Aviation Psychology* (pp. 350–355). Columbus: Ohio State University.

Predmore, S. (1993). *The dynamics of group performance: A multiple case study of effective and ineffective flight crew performance.* Unpublished doctoral dissertation, University of Texas at Austin.

Reason, J. T. (1990). *Human error.* Cambridge, England: Cambridge University Press.

Proceedings of the International Aerospace Congress IAC '94, (Edited by M. Liberzon), Kanki, B.J., Communication research in aviation and space operations: Symptoms and strategies of crew coordination, *pp. 160-165, (1995)* with kind permission from International Aerospace Congress, Moscow, Russia.

COMMUNICATION RESEARCH IN AVIATION AND SPACE OPERATIONS: SYMPTOMS AND STRATEGIES OF CREW COORDINATION

B. G. Kanki
Research Psychologist,
NASA Ames Research Center, MS 262-4
Moffett Field, CA 94035-1000 USA

Abstract. NASA Ames Crew Factors researchers have been developing a model of effective crew coordination in order to better understand the sources of team performance breakdowns, and to develop effective solutions and interventions. Because communication is a primary mechanism by which information is received and transmitted, and because it is observable behavior, we focus on these group processes in order to identify patterns of communication that distinguish effective from ineffective crew performance. Our research objective is to enhance communication practices through (1) the training of specific communication skills, and (2) the design of equipment, tasks, procedures, and teams that optimize smooth, unambiguous communication processes.

Two examples of communication research are described; one in aviation, and one in space operations. The first example is a simulation study which investigates the effect of flightdeck automation on crew coordination and communication (contrasting crew performance in the DC-9 vs. MD-88). While overall performance was not significantly different across the two levels of automation, we identified verbal and non-verbal communication patterns which point to differences in coordination strategies, and possible shifts in traditional workroles. The second example is a case study of a recent commercial launch incident. In this analysis, the organizational, procedural and training factors contributing to the breakdown in crew communication and coordination are discussed.

Key words: human factors, crew coordination, communication patterns, Crew Resource Management, team performance, accident and incident analysis, automation effects, information transfer, coordination strategies.

INTRODUCTION

The day-to-day operators of today's aerospace systems work under increasing pressures to accomplish more with less. They work in complex, high-risk operational systems in which incidents and accidents have far-reaching and costly consequences. For these and other reasons, there is concern that the safety net formerly built upon redundant systems and abundant resources may become overburdened. Although we know that human ingenuity can overcome great odds, human nature can also fail in unpredictable ways.

Over the last 20 years, 60-70% of aviation accidents and incidents have implicated human error rather than hardware or environmental factors alone (Lautman & Gallimore, 1987). Evidence provided by the accident investigations of the National Transportation Safety Board and reports from the Aviation Safety Reporting System (incident data) have led researchers to study this class of errors more thoroughly. Generally speaking, these errors appear not to be due to a lack of individual, technical competencies, but to the failure of teams to utilize readily available resources or information in a timely fashion. These insights helped motivate a training revolution in the aviation industry called Crew Resource Management (CRM) and its principles have been widely incorporated into both civil transport and military training programs in the U.S. and many other countries (see Wiener, Kanki & Helmreich, 1993).

Communication is a cornerstone in CRM training since crew coordination and resource management largely depends upon successful information transfer both within flightcrews, and between flightcrews and the ground control teams that support them. The research I shall describe takes its roots in CRM history, which drew our attention to communication processes as a means to discover symptoms of crew coordination problems, as well as strategies of effective crew management. On the one hand, communication is often the means or the tool by which team members manage their resources, solve problems, maintain situational awareness and procedural discipline, and establish a constructive interpersonal climate. Conversely, poor communications may result in the lack of planning and resource management, loss of vigilance and situational awareness, non-standard practices, and lack of leadership. These kinds of behaviors are symptoms of crew problems and are often implicated in accidents and incidents.

A prime objective of our research is to make recommendations that enhance team coordination and communication. In order to ensure operational relevance, we must interpret our research findings within the context of relevant task and environmental conditions, role and procedural constraints, and the normal real-time parameters of flight operations. Figure 1 illustrates multiple categories of input factors that can ultimately affect team performance outcomes such as safety and effectiveness. The process variables in the center box represent the type of behaviors we study (e.g., communication patterns, management styles, problem solving strategies, etc.) in order to better understand the ways in which crews effectively, or ineffectively handle problem situations.

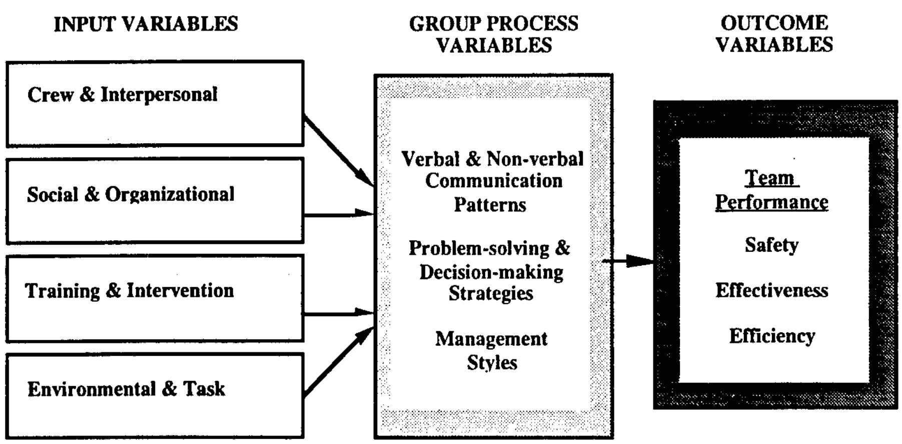


Figure 1: Conceptual Model for Team Performance (adapted from McGrath, 1984)

STUDY 1: AVIATION OPERATIONS

Increasingly automated aircraft has raised the issue of human-centered flightdeck design (see Billings, 1991), as well as the need to better understand the effects of flightdeck automation on crew coordination and communication. The following study examines these issues by means of a full-mission simulation study contrasting crew performance in the MD-88 vs. DC-9-30 (MD-88 is the glass cockpit derivative of the DC-9 series aircraft.). Using the conceptual model described above, this study designates "level of automation" as the task input variable and defines "team performance outcomes" in terms of observer ratings, errors and self-reported workload measures (see Figure 2). The group processes investigated were verbal and non-verbal communication patterns.

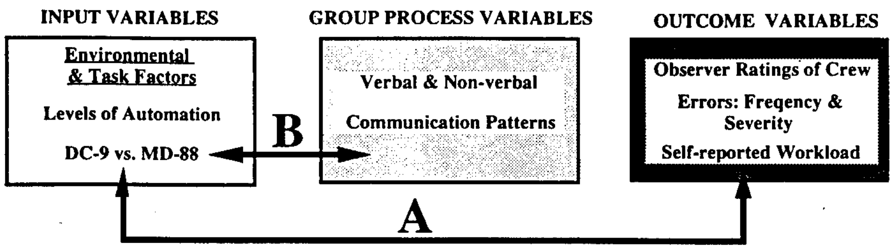


Figure 2: Conceptual Model of the Automation Simulation Study

Overall performance was not found to be significantly different across the two levels of automation (relationship "A" in Figure 2). However, MD-88 crews reported slightly higher workload, thus raising the question of whether MD-88 crews were working harder in order to achieve the same level of performance. We next investigated the "B" relationship, the effect of automation on verbal and non-verbal communication patterns, in order to understand the way in which crews in each aircraft type responded to the flight task demands.

Methodology. Communications of 12 DC-9 crews and 10 MD-88 crews were analyzed from the high fidelity full-mission simulation. Each 2-person crew consisted of a captain (CA) and first officer (FO) who were active line pilots from a single airline. Each crew flew a highly realistic flight scenario in which there was a normal and abnormal phase. The abnormal phase consisted of compounding factors including a constant speed drive overheat, weather deterioration, an unpublished missed approach and holding pattern, and deviation to an alternate airport. (Details on this study may be found in Wiener, Chidester, Kanki, Palmer, Curry, & Gregorich, 1991.)

Following the methodology developed in earlier studies (e.g., Kanki, Greaud-Folk & Irwin, 1991), communications were transcribed from the 22 videotaped simulation flights from "clear to push" to touchdown, an average of 80 minutes. Communications were organized into speech units, in real-time sequence, with start and end times noted. After all speech units were coded into one of 14 speech act categories, they were simplified into 7 categories and re-grouped into 2-part sequences consisting of initiation speech (4 types) followed by response speech (3 types). Because either CA or FO could initiate a sequence, there were two possible speaker directions for each initiation-response pair; CA -> FO or FO -> CA. Therefore, for each speaker direction, a communication matrix could be constructed in which the frequencies and proportions of each speech type and sequence could be tabulated. As shown in Figure 3, the four initiation speech categories were commands, observations, questions and dysfluencies. The three response categories were replies, simple acknowledgements and no response.

CA --> FO	Replies	Acknowledge-ments	No Response	Row %
Commands				
Observations				
Questions				
Dysfluencies				
Column %				Total CA-> FO Speech Sequences

Figure 3: Communication Matrix of Speech Sequences in the Captain to First Officer Direction

From these simple frequency matrices, it was found that MD-88 crews produced more total speech, and in particular, captain questions appeared to stand out as a differentiating characteristic. Therefore, questions were further coded in terms of (1) their function (seek information vs. verify information), (2) their topic (system vs. navigation vs. procedure), and (3) the response they received (confirm/disconfirm answers vs. information answers vs. no answer)

Results included the following: Compared to DC-9 crews, (1) MD-88 crews had more information-seeking questions and answers (no differences with respect to verification questions), (2) MD-88 crews asked more questions on navigation and system topics (no differences on procedure questions), and (3) MD-88 crews had more questions left unanswered (no difference with respect to confirm/disconfirm answers). Finally, most of the findings were stronger during the abnormal phase. These results led us to ask whether the automation was creating a situation in which information access (on both navigation and systems topics) was less direct or less clear in the MD-88.

A follow-up study focusing on non-verbal activity was also conducted with this dataset. Because results were typically stronger during abnormal phase, non-verbal control actions were coded during the 10 minutes following missed approach only. During this period, CA was the pilot flying while FO was the pilot not flying. Three kinds of behaviors were coded: navigation actions, systems actions, and procedure actions. Obviously, much of the MD-88 activity centered around computer entry. Results from these analyses revealed that traditional work roles were maintained by the DC-9 crews (i.e., CA exhibited more systems actions and FO exhibited more navigation actions), but the MD-88 crews showed very different patterns. Specifically, CA and FO showed about the same amount of systems actions, and the CA showed more navigation actions than FO in spite of the fact that he was also the pilot flying. Thus, in addition to the information access issue, automation appears to affect work management resulting in a shift away from traditional work roles. While this may be a constructive shift (part of a successful crew strategy), any changes in practices should be thoroughly understood so that training and procedures are compatible, and that changes are standardized for all pilots. (Full details are in Kanki, Veinott, Irwin, Jobe, & Wiener, in prep.)

STUDY 2: SPACE OPERATIONS

A second example of communication research is a case study in the space operations domain; namely an analysis of team coordination and communication during a commercial launch incident. In the final few minutes of countdown, confusion in the control room resulted in the continuation of the launch of an expendable launch vehicle in spite of range safety's call for abort. "... Misunderstanding of communication channel assignments... lack of adequate coordination... and inadequate rehearsal of off-nominal... events led to the launch team's failure to properly abort the mission or rescind the abort call... " (McKenna, 1993, June 21, p. 62). Although the launch and subsequent deployment of satellites were totally successful, and there were no injuries or damages, the potential for costly consequences was great. The following section of transcript (Figure 4) clearly illustrates the breakdown in information transfer during the critical minute before launch. In short, the call for "Abort" by the test director (TD) was never successfully communicated to the test conductor (TC), communicator (COM) and B52 pilot.

All speech transmissions from T-3:54 to T+1:31 were coded by time, speaker's organization and channel assign-ments, by the response received, and use of standard protocol. A summary of these data include the following: (1) Of the total 140 speaker transmissions on 4 channels, 53% were communications within Organization A, 40% were within Organizations B and C, and only 7% transmissions crossed the organizational boundaries between A and (B&C). (2) Cross-organization communication took place on channel 4 only and appeared to be successful only in the direction of range control (RCO) to the communicator to the aircraft (COM). (3) There was no cross-organizational communication among the principal managers (e.g., test director and test conductor, TD and TC), nor were there successful cross-organizational communications during the critical one minute to launch shown in Figure 4.

TIME	CH 1	CH 4	UHF CH12
-T :56	TD: Abort		
-T :44		TC -> COM: Abort	
-T :34	RCO <-> RSO		COM -> B52: Abort
-T :27	RSO <-> RSS		B52 -> COM: Understand
-T :23		TC -> COM: Negative	
-T :22			COM -> B52: Negative on Abort...Keep going
-T :18	RSO <-> RSS		
-T :04	RSO <-> RCO		B52 -> COM: Go for launch
-T :00		RCO: Abort	
+T :01			B52 -> COM: Away
+T :05		TC: Away	

Organization/Role	Channel 10	1	4	12	Organization/Role	Channel 10	1	4	12
A / TD test director	X	X	-	-	B / TC test conductor	-	-	X	X
A / RCO range control	X	X	X	-	C / COM communicator	-	-	X	X
A / RSO range safety	X	X	-	-	C / B52 pilot	-	-	-	X
A / RSS range safety support	X	-	-	-					

Figure 4: Launch Incident Transcript from -T :56 to +T :05

To summarize, it is clear that there was inadequate cross-organizational communication, and this was highly related to the fact that authority and procedures were unclear, and the use of communication channels was poorly assigned. In addition, there was an inconsistent use of protocol across Organizations A, B & C. While Organizations B and C were more formal in their use of callsigns and standard radio phraseology and acknowledgements, Organization A was informal, and made inconsistent use of callsigns and personal names (for fuller description, see Kanki, 1993).

In conclusion, our research in both aviation and space operational domains has been directed toward identifying communication patterns related to performance differences. Understanding the conditions under which patterns occur, their relationship to organizations and work roles, as well as optimal timing help us to distinguish between patterns which are symptoms of problems and those which are strategies for successful problem solution. When such communication practices are distinguished, our second goal is to develop means of enhancing team performance through more effective communication practices. Recommendations are made in the areas of team training and for the design of tasks, teams and procedures that optimize crew communication and coordination.

REFERENCES

Billings, C. E. (1991). *Human-centered aircraft automation: A concept and guidelines.* (NASA Technical Memorandum 103885). Moffett Field, CA: NASA Ames Research Center.

Kanki, B. G. (1993) *A Crew Factors Perspective on the Pegasus/SCD-1 Launch Operation Incident.* included as Appendix E (NASA Ames Research Center Letter on Crew Coordination and Communication) Special Investigation Report / NTSB SIR-93/02, Commercial Space Launch Incident.

Kanki, B. G., Greaud, V. A. & Irwin, C. M. (1991). Communication variations and aircrew performance. *International Journal of Aviation Psychology. 1,* 2, 149-162.

Kanki, B. G., Veinott, E. S., Irwin, C. M., Jobe, K. & Wiener, E. L. (in prep). *The impact of cockpit automation on crew coordination and communication: II. Analyses of speech and non-verbal activity across pilot roles and phases of flight.* (NASA Technical Memorandum). Moffett Field, CA: NASA-Ames Research Center.

Lautman, L. G. & Gallimore, P. L. (1987, April-June). Control of the crew caused accident: Results of a 12-operator survey. In *Airliner:* Magazine of the Boeing Commercial Airplane Co., pp.1-6.

McGrath, J. E. (1984). *Groups: Interaction and performance.* Englewood Cliffs, N.J.: Prentice-Hall.

McKenna, J. T. (1993, June 21). NTSB cites confusion in Pegasus launch room. *Aviation Week & Space Technology*, p. 62.

Wiener, E. L., Chidester, T. R., Kanki, B. G., Palmer, E. A., Curry, R. E., and Gregorich, S. E. (1991). *The impact of cockpit automation on crew coordination and communication: I Overview, LOFT evaluations, error severity, and questionnaire data.* (NASA Contractor Report No. 177587). Moffett Field, CA: NASA Ames Research Center.

Wiener, E. L., Kanki B. G., & Helmreich, R. L., (Eds.) (1993). *Cockpit Resource Management.* San Diego: Academic Press.

Collaboration in Controller-Pilot Communication

Daniel Morrow
Decision Systems

This research was supported by NASA-Ames Research Center under a contract to Sterling Software (#2000-00342240), and by the National Institutes of Health (NIAAA #AA07035). An earlier version of this paper was presented as part of the Aerospace Human Factors Research Division Seminar Series, October, 1993.

I. Overview

Like other forms of dialogue, ATC communication is an act of collaboration between two or more people. Collaboration progresses more or less smoothly depending on speaker and listener strategies. For example, we have found that the way controllers organize and deliver messages influences how easily pilots understand these messages, which in turn determines how much time and effort is needed to successfully complete the transaction. In this talk, I will introduce a collaborative framework for investigating controller-pilot communication and then describe a set of studies that investigate ATC communication from two complementary directions. First, we focused on the impact of ATC message factors (e.g., length, speech rate) on the cognitive processes involved in ATC communication. Second, we examined pilot factors that influence the amount of cognitive resources available for these communication processes.

These studies also illustrate how the collaborate framework can help analyze the impact of proposed visual data link systems on ATC communication. Examining the joint effects of communication medium, message factors, and pilot/controller factors on performance should help improve air safety and communication efficiency. Increased efficiency is important for meeting the growing demands on the National Air System.

II. Collaboration in ATC Communication

We analyzed ATC communication by adapting a general model of collaboration (Clark & Schaefer, 1989). Similar approaches have been used to analyze many kinds of dialogue, including human-computer interaction and automated voice systems (Karis & Dobroth, 1991). According to this view, transactions between speakers and listeners involve three collaborative phases, which participants accomplish by using several types of speech acts (Figure 1, see Morrow, Rodvold, & Lee, 1994, for more detail). These phases provide several points of comparison between voice and data link. These comparisons show how the same collaborative function can be accomplished in different ATC communication systems.

A. Collaborative phases

Initiate transaction. The speaker first attracts the listener's attention in order to initiate or open the transaction. Typically, air-ground ATC transactions begin with aircraft or facility call signs. Stress and intonation can also be critical for attracting attention in voice communication. In data link systems, message announcement chimes and visual alerts on the data link display may be used to start the transaction (ATA, 1992).

Present message. The speaker presents the message once the transaction is initiated. Commands, reports, or requests are presented by voice or computer. Understanding these messages requires cognitive processes such as word recognition, parsing, and updating a mental model of the current situation with the new information.

Accept message. Communication requires more than simply presenting and understanding messages-- The speaker and listener must collaborate in order to accept the message as mutually understood. This is particularly important in ATC communication, which requires accurate understanding at a detailed, utterance by utterance level (Morrow et al., in press). In the current voice environment, acceptance rests upon pilot acknowledgments with readback and callsign. The controller in turn must "hear back", or monitor the acknowledgement to ensure that the message was understood. Several acceptance procedures have been proposed for data link systems, including a digital accept/reject response and an intracrew readback (ATA, 1992). In either voice or data link systems, the controller and pilot must agree before continuing that they understand the message and share a mental model of how the intended actions will change the flight situation. In ATC parlance, this phase is called "closing the communication loop".

B. Cognitive Resources and collaboration

The cognitive processes underlying collaboration depend on speaker and listener cognitive resources, which are limited in quantity. For example, noticing, understanding, and accepting messages require selective attention and working memory capacity. The constraints imposed by limited cognitive resources is often illustrated by a diagram of the flow of information through a series of processing steps. However, this individual-centered approach must be expanded to include the shared cognitive resources required by collaborative effort-- the speaker and listener resources needed to achieve mutual understanding (Figure 2). For example, pilot responses such as partial acknowledgments can increase demands on controller working memory by forcing controllers to repeat the message and to continue monitoring the transaction. This increases the overall shared resources needed to close the transaction. The notion of collaborative effort has been useful for analyzing telephone dialogue (Clark & Schaefer, 1989) and crew coordination (Kanki, Lozito, & Foushee, 1989). Our studies have examined trade-offs of individual and collaborative effort in controller-pilot communication.

C. Factors influencing available cognitive resources

The success of communication depends on available individual and collaborative cognitive resources as well as the demands imposed on these resources by communication. The amount of resources available for communication depends on short term factors such as fatigue and distraction and longer term factors such as age and experience (Morrow & Rodvold, in press). The second set of studies that I'll describe examined the influence of pilot age on ATC communication.

D. Communication problems

Several types of problems tend to arise when available resources do not meet task demands during each collaborative phase.

Initiate transaction. Problems during initiation often relate to attention failures. For example, aircraft crew may not hear an ATC call because they talk over the message. They may also misunderstand the intended addressee, creating callsign confusions. Initiation problems have been an important impetus for discrete address data link systems (ATA, 1992).

Present message. A transaction may be successfully initiated, but message is misunderstood or misremembered because message presentation overloads working memory. The visual data link medium should reduce memory problems, although message complexity could amplify problems associated with poor data link menu or interface design.

Accept message. Problems during the acceptance phase often relate to failure to follow acknowledgment procedures. For example, controllers and pilots may fail to explicitly close transactions because of missing acknowledgments (Morrow et al, 1994). There is also a concern that acknowlegment delays may disrupt data link transaction organization (ATA, 1992). The present studies focused on problems related to presenting and accepting messages.

III. Studies of ATC Communication

So far, I've sketched collaborative phases in ATC transactions, cognitive processes and resources involved in each phase, and possible communication problems. Our studies examined how pilot communication problems arise when complex ATC messages tax cognitive resources. We focused on message complexity because it is a concern in the National Air System (Billings & Cheany, 1981; Cardosi, 1993; Morrow, Lee, & Rodvold, 1993) and because manipulating complexity helps to map relations between demands imposed by message and medium factors, available cognitive resources, and communication problems. It also helps illustrate trade offs between individual and collaborative effort. For example, why do controllers produce complex messages, and what are the consequences of this strategy?

We started with a field study in order to generate hypotheses about problems related to message complexity. This was followed by laboratory studies that tested some of the hypotheses. These studies were conducted at NASA-Ames Research Center. A second set of laboratory studies (conducted at Stanford Medical Center) compared the performance of older and younger pilots on ATC communication tasks. According to cognitive aging theory (Salthouse, 1985), older pilots should have fewer cognitive resources than their younger counterparts. Therefore, we can indirectly examine the role of resources in ATC communication by means of this age comparison. These studies also relate back to the concept of collaborative effort-- pilots or controllers with fewer cognitive resources may be more likely to use strategies that minimize their effort at the expense of the other person.

A. Message complexity and ATC communication: Field study

Introduction. As a first step, Michelle Rodvold and I analyzed communication between controllers and pilots during daily operations in the terminal environment. Before this study, there was little information about routine ATC communication other than from incident/accident analyses. Therefore, we wanted to collect base rate information on the frequency of problems related to message complexity. The study would also provide a snapshot of collaborative processes in routine ATC communication.

Method. We collected 42 hours of taped communication (almost 8000 transactions) from four of the busiest TRACONs in the United States. Communication was transcribed and coded utterance by utterance for speech acts and topics (Figure 3; see Morrow, et al, 1993 for more detail). We also focused on nonroutine transactions, where the pilot or controller interrupts routine information transfer in order to clarify miscommunication, for more elaborate analysis of collaborative strategies (Morrow et al., 1994).

Results. First of all, longer messages (with 3 or more information units such as commands and reports) occurred in 5-20% of transactions, depending on the TRACON sample. More complex messages were associated with pilot comprehension and memory problems. For example, readback errors increased with message length (Figure 4). A similar pattern has been found for transactions in the enroute environment (Cardosi, 1993). Analysis of readback errors in our sample suggested that long messages taxed working memory. For example, incorrect digits in pilot readbacks often came from other numbers in the same message (intrusion errors), suggesting the error was due to interference.

Message complexity also disrupted the acceptance phase of transactions. Pilot acknowledgements were more streamlined after longer ATC messages, since the number of partial readbacks increased with message length. Thus, after delivering long messages, controllers are more likely to have to get back on the radio and request full acknowledgment.

Message length also influenced the way in which pilots read back the message (Figure 5). Pilots coped with longer messages by using strategies that minimized memory load (in addition to reading back less information). After shorter messages, they tended to say their call sign before the readback, as recommended by the Airmen's Information Manual. After longer messages, they tended to say the call sign after the readback. While this strategy may minimize memory load (repeat the new information first), it complicates the hearback because the controller has to wait until the end of the readback to make sure that the correct pilot responded.

Pilots also tended to repeat short messages verbatim, with commands in the same order as in the message. With longer messages, they tended to paraphrase and to repeat commands in a different order. These findings are not surprising in light of laboratory studies showing that verbatim memory tends to drop off after complex messages and/or long retention intervals (Anderson, 1980). But they also make an important point about collaboration in the ATC environment-- Readbacks after longer ATC messages tend to be less similar to the message in terms of terminology and information order, which may complicate the hearback part of the communication loop. Longer messages also tend to increase the number of communication problems, which lead to nonroutine transactions in which the communication is clarified (Cardosi, 1993). These transactions are longer than routine transactions and less efficient because the extra turns are devoted to clarifying old information rather than presenting new information. ATC language is also less standard and more complex in nonroutine transactions, which may lead to further confusion (Morrow et al., 1994).

In summary, our field analyses suggest a trade-off between individual and collaborative effort (Figure 6). Controllers sometimes deliver long, complex messages, perhaps to reduce turn-taking time and thus their own cognitive effort. These messages may overload pilots' cognitive resources so that the pilots misunderstand the message, request clarification, or adopt acknowledgement strategies that ease demands on memory. Any of these consequences can increase the difficulty of accepting the message and closing the transaction, resulting ultimately in greater collaborative effort.

B. Message complexity and ATC communication: Part-task simulation study

Introduction. We conducted a part-task simulation study to provide more conclusive evidence for the impact of message complexity on communication (see Morrow & Rodvold, 1993 for more detail). This study was conducted at NASA in collaboration with Michelle Rodvold, Sandy Lozito, Alison McGann, and Kevin Corker. With the help of several controllers and pilots, we created flight scenarios in which pilots were vectored by ATC in enroute and terminal environments. ATC messages were delivered in two ways: One long message with 4 commands (e.g., heading, altitude, speed, frequency) or a pair of short messages with 2 commands each. By delivering the same content in different ways, we could examine message length independent of content. Because controllers delivering two messages to the same aircraft would want to minimize communication time, we decided on a brief interval between delivering each message, roughly 10 sec.

Based on the earlier field results, we expected pilots to have more communication problems when confronted with the longer messages-- more requests for clarification and readback errors. However, short messages may also create problems. Because the second message of the sequence is delivered so quickly after the first, it may interfere with the pilot's response to the earlier message, resulting in delayed requests to clarify this message. The impact of these message factors on data link as well as voice communication was examined in a parallel study. Some data link findings will be mentioned at the end of the talk.

Method. The part-task laboratory consisted of (a) Workstation simulating an ATC radar station; (b) Workstation simulating a flight deck display; (c) Macintosh computer that presented the pre-recorded ATC messages. These computers were networked so that the controller could track the subject's aircraft on radar and control delivery of the messages to the flight deck display. The controller and pilot were also linked by a telephone-radio system. The flight scenarios imposed experimental control but also allowed for interactive communication. Scripted ATC messages were recorded, digitized, and sent by the controller to the pilot via computer. Pilots responded to these messages as they flew, and the controller was present in order to handle radio clarifications. Sixteen air carrier pilots participated in the study.

Results. Figure 7 shows that pilots were more likely to misunderstand the controller when too much information was presented in one message. They were more likely to ask for clarification after longer messages than after the two short messages combined. They also made more readback errors after longer messages (18% after long messages, 8% after short messages).

Problems associated with short messages differed from those after long messages (Figure 8). Pilots initially understood the first short message-- In most problems, they had first correctly read back the commands immediately after the first short message. However, they often forgot all or part of the first message by the time the second occurred-- Most of the delayed problems were requests for repeat or were incorrect requests for confirmation. These incorrect requests often contained intrusions, with one or more digits from the second message. Thus, pilots usually understood the first short message but then forgot part of it either because the second one created interference or delayed response to the first.

We recently conducted a second part-task study in order to systematically examine the impact of message interval on voice and data link communication (see Morrow, 1994 for more detail). While message interval was fixed in the first study, it was manipulated in this follow-up experiment. The second message was delivered either 5 sec or one min after the readback of the first message. In addition to voice and data link communication, we examined a mixed voice/data link environment where a voice ATC message was followed by a data link message, or viz versa. The mixed environment was examined because parts of the ATC system will likely resemble this hybrid when data link is introduced into the current environment (ATA, 1992).

Figure 9 shows that more voice communication problems (e.g., requests for repeat) occurred when voice messages were presented with short rather than long intervals (the difference between voice-only and mixed environments was not significant). Unlike the previous study, these problems usually related to the second rather than the first message of the sequence. Pilots delayed responding to the second message in order to complete their response to the first, and thus sometimes had to clarify the second message. Nonetheless, both part-task studies show that communication problems can arise from time pressure imposed by a rapid sequence of ATC messages.

The findings from these laboratory studies converge with the field results to show trade-offs between individual and collaborative effort in ATC communication-- Controllers may try to save time and effort by delivering too much information in one message or by delivering messages in quick succession. However, these strategies may increase collaborative effort and reduce communication efficiency by creating pilot comprehension or memory problems.

C. ATC Message factors and available cognitive resources: Age and practice

Introduction. So far we've examined comprehension and memory problems in ATC communication by investigating the influence of message delivery on ATC communication. The final studies examined how communication depends on the cognitive resources that pilots have available for meeting the demands imposed by communication. These studies were conducted at Stanford Medical Center in collaboration with Von Leirer, Jerry Yesavage, Joy Taylor, and Nancy Dolhert.

Because aging is often associated with a gradual decline in cognitive resources such as working memory capacity (Salthouse, 1985), comparing older and younger pilots provides a way to analyze the impact of cognitive resources on ATC communication. While older pilots may usually perform as well as younger pilots (e.g., because of selection effects, compensation of experience for age declines), age differences may arise for difficult ATC tasks. Therefore, we examined older and younger pilot performance on scenarios similar to our earlier studies (Morrow, Yesavage, Leirer, & Tinklenberg, 1993). The earlier studies suggest that long messages impose heavy demands on working memory. Such messages may particularly penalize older pilots if they have fewer cognitive resources to devote to communication, especially because they have to divide attention across other flying tasks while communicating. We also examined if practice on the communication tasks differentially improved older pilot performance. This might occur if older pilots were relatively unfamiliar with complex ATC communication tasks to begin with. In addition to providing a window on cognitive processes in ATC communication, findings about aging and pilot performance may have implications for the Age 60 retirement rule for Part 121 pilots in the United States.

Method. Fifteen older (Mean age= 38 years) and 16 younger (mean age=26) private license pilots flew a light single engine aircraft simulator with computer-generated out- the-window visuals. Older and younger pilots did not differ in terms of health, education, or flying experience. As in the part-task studies, ATC messages were pre-recorded and the scenarios involved vectoring in a terminal environment. Pilots flew 12 flights and performance was averaged across each set of 3 flights. Therefore, we examined ATC communication (readback and execution errors) and flying performance for older and younger pilots over the 4 flight sets.

Results. Figure 10 shows that older pilots made more readback and execution errors than younger pilots. Practice improved performance for both age groups but did not reduce age differences. Readback errors included intrusions from other parts of the message, providing further evidence that long messages can overload working memory. Finally, age differences were minimal for flying performance that did not depend on communication (e.g., deviation from center line on take off and landing). Thus, the older pilots generally flew as well as the younger pilots, but they had more difficulty with the heavy memory demands imposed by the communication task.

D. ATC Message factors and cognitive resources: Age, message length, and speech rate

Introduction. We also examined the joint effects of message complexity and pilot age on communication (Taylor et al, 1994). Older and younger pilots in this study responded to messages varying in length and speech rate. "Speedfeed" is a frequent pilot complaint (Morrison & Wright, 1989), and laboratory studies show that recall declines as speech rate increases, particularly for older adults (Stine, Wingfield, & Poon, 1986). Therefore, faster as well as longer ATC messages should increase demands on cognitive resources and produce communication problems. Older pilots may be particularly vulnerable to these messages because of age-related resource declines. On the other hand, speech rate effects are reduced for more meaningful or predictable texts (Stine, et al., 1986). Thus, older pilots may be able to compensate for reduced cognitive resources by relying on knowledge of ATC message structure.

To more directly test if the impact of message complexity on communication is mediated by working memory limits, individual differences in working memory capacity were measured by the WAIS-R digit span test. Correlations between span scores and communication errors would provide more direct evidence that the errors reflect working memory limits.

Method. Fifteen older (Mean age= 61) and 15 younger (mean age=28) pilots with instrument ratings flew in the same simulator as in the previous study. Half of the messages in each scenario contained 3 commands and half contained 4 commands. The messages were recorded by a controller at a typical speech rate (235 wpm). For both long and short messages, half were time-compressed (while minimizing pitch distortion) to produce a rate that was 50% faster than the normal version.

Results. Figure 11 shows that older pilots again made more message execution errors. In addition, longer messages (long=45%, short=23%) and messages presented at the faster rate (fast=37%, normal=31%) produced more errors. Age and message complexity had additive effects on communication. Thus, age differences were not magnified by difficult messages. Notably, pilots with higher span scores produced fewer errors ($\underline{r}$=-.47), providing some evidence that message factors influenced performance through their effects on working memory.

Discussion. These studies show that aging can influence pilot performance of very demanding communication tasks. However, the studies involved noncommercial pilots. Using pilots with relatively low levels of experience may overestimate age effects in pilot performance. In fact, we have found some evidence that expertise reduces age differences in a laboratory readback task (Morrow, Leirer, Fitzsimmons, & Altieri, 1994). Nonetheless, the pilot age studies suggest that individual differences in performance may be useful for studying the role of cognitive resources in ATC communication.

E. ATC communication medium

I'll conclude by summarizing some data link findings from our part-task studies, which show how ATC message factors can interact with the communication medium to influence pilot-controller communication. The first part task study found that while message length had a large impact on voice communication, it had little effect on data link acknowledgement time or requests for clarification (McGann, Lozito, & Corker, submitted). Because of the relatively permanent visual medium, complex ATC messages appear to impose few demands on pilot working memory in data link systems. Of course, message complexity could become an issue if the menus and interface in data link systems impose demands on pilot working memory.

Data link communication was not immune to problems in our studies. For example, the short interval between messages slowed data link as well as voice communication in both part-task studies. The fact that the dynamics of message delivery can influence data link as well as voice communication reinforces concerns about introducing data link into busy terminal environments (ATA, 1992). These kinds of studies should help identify collaborative strategies for coping with dynamic communication in terminal environments, whether voice or data link is used.

IV. Conclusions

In summary, we have investigated how pilots and controllers collaborate to ensure mutual understanding in busy environments. We focused on readback/hearback procedures because they are essential to safe and efficient communication. The effectiveness of these procedures depends on the demands on pilot and controller cognitive resources imposed by ATC message and medium factors, as well as on available cognitive resources. For example, our field and laboratory studies show that problems arise in voice environments when pilot working memory is overloaded by long, fast messages, or by shorter messages presented in quick succession. Additional time and effort is then needed in order to clarify the problem and accept the message as mutually understand, creating increased collaborative effort. Pilots can also increase controller workload by not following acknowledge procedures. For example, missing or partial acknowledgments may force the controller to repeat the message.

Our studies of pilot age and ATC communication suggest that complex or nonstandard communication may particularly tax older (or fatigued) pilots and controllers, who may have fewer resources to devote to the task. Therefore, they are more likely to use short cuts that reduce their own effort at the expense of collaborative effort. Some collaborative problems may be alleviated by a change in communication medium, while others remain. With visual data link, pilots may be able to easily handle long ATC messages but still have problems with a series of messages delivered in quick succession.

These findings suggest several recommendations for improving voice communication procedures, such as the optimal length and timing of ATC messages in the terminal environment. The collaborative framework also has training implications. Pilot and controller training should stress the importance of trade-offs between individual and collaborative effort--When individuals reduce their own effort at the expense of other participants, everyone's workload tends to increase and accuracy and efficiency suffers. The concept of collaboration also has broader, more organizational implications. Pilots and controllers are more likely to collaborate during air-ground communication if they understand each other's responsibilities and constraints. Collaboration must be fostered rather than inhibited by organizational boundaries (SAE ARD #50045, in preparation).

References

Air Transport Association, (1992). Research agenda for data link: Proposed for inclusion in the ATC/Flight Deck Integration section of the National Plan for Aviation Human Factors.

Anderson, J. (1980). Cognitive psychology and its implications. San Francisco: W.H. Freeman.

Billings, C., & Cheaney, E. (1981). Information transfer problems in the aviation system. NASA technical paper 1875.

Cardosi, K. (1993). An analysis of en route pilot-controller voice communications. U.S. Department of Transportation, Federal Aviation Administration.

Clark, H., & Schaefer, E. (1989). Contributing to discourse. Cognitive Science, 13, 259-294.

Karis, D., & Dobroth, K. (1991). Automating services with speech recognition over the public switched telephone network: Human factors considerations. IEEE journal on selected areas in communications, 9, 574-585.

Kanki, B., Lozito, S., & Foushee, C. (1989). Communication indices of crew coordination. Aviation, Space, and Environmental Medicine, 56-60.

McGann, A., Lozito, S., & Corker, K. (submitted). An experimental evaluation of data link textual formats.

Morrison, R., & Wright, R. (1989, April). ATC control and communications problems: An overview of recent ASRS data. In R.S. Jensen (Ed.), Proceedings of the Fifth Symposium on Aviation Psychology, 2, Columbus, Ohio.

Morrow, D. (1994). Part-task study of pilot-controller communication: voice communication problem results. Subcontract Report 0342-6-240. Sterling Software, Inc. Foster City, CA.

Morrow, D., Lee, A., & Rodvold, M. (1993). Analyzing problems in routine controller-pilot communication. International Journal of Aviation Psychology. 3, 285-302.

Morrow, D., Yesavage, J., Leirer, V, & Tinklenberg, J. (1993). The influence of aging and practice on piloting tasks. Experimental Aging Research., 19, 53-70.

Morrow, D., Leirer, V., Fitzsimmons, C., & Altieri, P. (1994). When expertise reduces age differences in performance. Psychology and Aging, 9, 134-148.

Morrow, D., & Rodvold, M. (1993) The influence of ATC message length and timing on pilot communication. NASA contractor report 177621.

Morrow, D., & Rodvold, M. (in press). Communication issues in Air Traffic Control. in M. Smolensky & E. Stein (Eds.), Human factors in Air Traffic Control. Academic Press.

Morrow, D., Rodvold, M., & Lee, A. (1994). Nonroutine transactions in controller-pilot communication. Discourse Processes, 17, 235-258.

Salthouse, T. (1985). Speed of behavior and its implications for cognition. In J. Birren & K. Schaie (Eds.), Handbook of the psychology of aging (2nd Ed, pp. 400-426), New York: Van Nostrand Reinhold.

Society of Automotive Engineers (in preparation). Issues impacting effective operational communication within the National Airspace System. SAE Aerospace Resource Document, ARD #50045.

Stine, E., Wingfield, A., & Poon. L. (1986). How much and how fast: Rapid processing of spoken language in later adulthood. Psychology and Aging, 1, 303-311.

Taylor, J., Yesavage, J., Morrow, D., Dolhert, N., & Poon, L. (1994). Effects of information load and speech rate on young and older aircraft pilots' ability to read back and execute ATC instructions. Journal of Gerontology: Psychological Sciences.

Proceedings of the Sixth International Symposium on Aviation Psychology, (Edited by R.S. Jensen), Predmore, S., Microcoding of cockpit communications in accident analyses: Crew coordination in the United Airlines flight 232 accident, pp. 350-355, (1991), with kind permission from Ohio State University.

Microcoding of Communications in Accident Investigation: Crew Coordination in United 811 and United 232

Steven C. Predmore[1]
NASA/University of Texas Crew Performance Project
The University of Texas at Austin

Two recent airline accidents have underscored the value of CRM to line operations, particularly under stressful, high workload conditions. On February 24, 1989, a United Airlines 747 lost a cargo door shortly after takeoff from Honolulu, leaving a gaping hole on the right side of the fuselage and the resulting loss of two engines. Although nine passengers were ejected from the aircraft and died as an immediate result of the explosion, the crew managed to make a successful emergency landing in Honolulu with no further loss of life (NTSB, 1990a). On July 19, 1989, a United Airlines DC-10 enroute from Denver to Chicago suffered a catastrophic failure of the #2 engine during cruise flight. The fragmentation and discharge of fan rotor parts in the tail section severely damaged all three hydraulic system lines resulting in the loss of all aircraft hydraulic operating systems. The crew was able to manage minimal flight control by the use of differential engine thrust and eventually performed an emergency landing at Sioux City Gateway Airport. Although there were 111 fatalities it has been widely accepted that the number of casualties would have been higher were it not for the performance of the flightcrew. The NTSB, in their analysis of the performance of the crew, summarized: ". . that under the circumstances the UAL flightcrew performance was highly commendable and greatly exceeded reasonable expectations." (NTSB, 1990b). In both cases the Captain cited training in Cockpit Resource Management as contributing significantly to the overall effectiveness of the crews. An analysis of the verbal behavior of each crew was undertaken to explore how catastrophic events impact upon the dynamics of crew interaction, and how the principles of CRM played out under stressful, high workload conditions contribute to successful crew performance. The case study approach we have taken here is viewed as complementary to the large scale observational methodologies reported elsewhere (Butler, in press; Clothier, in press).

MICROCODING OF COMMUNICATIONS

The verbal interactions of each crew were transcribed into our database from Cockpit Voice Recorder (CVR) transcripts obtained from the NTSB. In the case of United 811, the CVR captured crew interaction from about 8 minutes prior to the loss of the cargo door through the return and landing at Honolulu. The CVR in United 232 picked up crew communications from about 10 minutes after the loss of the engine through the crash landing at Sioux City. The current version of our coding scheme allows for the encoding of communications at a relatively detailed level termed, *thought units*. Thought units are utterances which deal with a single thought, intent or action. Each thought unit is classified as to speaker, target, time of onset, and speech form: Command-Advocacy, Inquiry, Incomplete-Interrupted, and Reply-Acknowledgment.

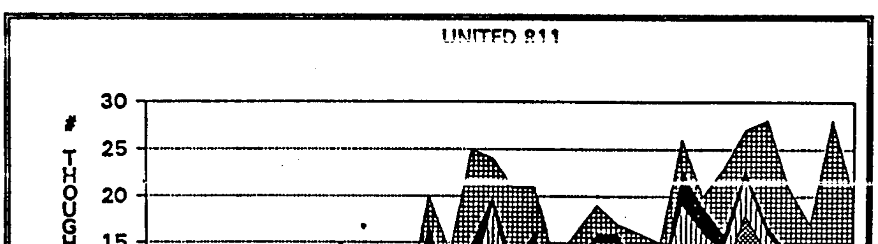


Figure 1. United 811: Communications broken down by speech form.

<u>Communication under stress.</u> Figure 1 is a stacked area chart which illustrates the cockpit and radio communications from United 811 broken down by speech form. The chart is two dimensional and represents the frequency of each speech form by the vertical thickness of its band at a given minute on the horizontal axis. This chart illustrates very clearly the impact of just such a catastrophic event on the level of crew interaction. The loss of the cargo door occurs at the end of minute 8, and we see an

immediate and sustained increase in the frequency of communications from minute 9 to the end of the CVR. Prior to the emergency, the crew averaged 5.13 thought units per minute, with a two minute period where there was no verbal communication at all. After the loss of the cargo door, the rate of communication rose to a mean of 18.85 thought units per minute.

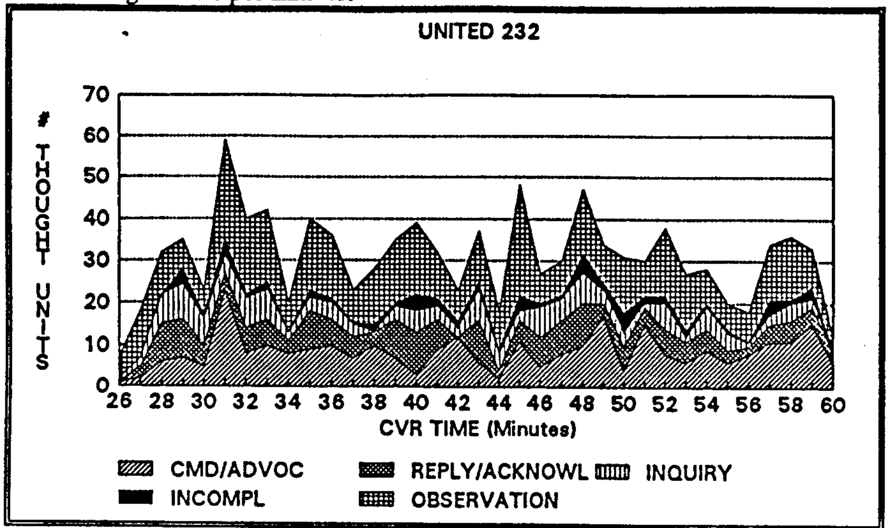


Figure 2. United 232: Communications broken down by speech form.

The same analysis was done for the United 232 accident and is presented in Figure 2.[2] At first glance, perhaps what is most striking about the communications chart for United 232 is the sheer volume of communication which occurs throughout this scenario. As many as 59 thought units are expressed in a single minute (Minute 31) with an overall mean rate of 30.86 thought units per minute. This rate is considerably higher than we have observed in other simulated and actual high workload flights. This significantly higher rate of communication is likely due to two factors. First, the cockpit of United 232 included a check pilot who had been seated in the cabin when the engine failure occurred, and who, as a fourth pilot in the cockpit, was responsible for 20% of the total communications. In fact, the only crewmember who contributed more than the Check Pilot in terms of communications was the Captain (49%). Second, fully 20% of the total number of communications were radio communications (originating from Sioux City Approach, United Maintenance, United Dispatch, etc.) as compared to United 811 where only 10.5% of communications originated outside of the cockpit.

One of the potential hazards of such a high rate of communication is the increased potential for information to be lost or misinterpreted through miscommunication. This is captured graphically in both charts by the black shaded band labeled 'INCOMPL' which denotes utterances which are incomplete, interrupted or unintelligible. A similar band appears in the chart of United 232. In both United 811 and United 232, the level of dysfluenies remains fairly constant and accounts for about 7% of the total communication. The pervasiveness of this effect seems to underscore the importance of educating crews as to the effects of stress in terms of the potential for lost or misinterpreted communications.

COMMUNICATIONS ACROSS TASKS

The scenarios encountered by the crews of United 232 and 811 are characterized by the need to accomplish a number of tasks concurrently within a limited amount of time under conditions of high stress. Yet, one of the more robust phenomena regarding performance under stressful conditions is the tendency for an individual's perceptual focus to narrow, resulting in a decreased ability to process multiple tasks. Under such conditions, teamwork is critical to effective performance, and involves distributing tasks across individual crewmembers, monitoring task processing, fully utilizing all available resources, and in the case of competition for limited resources, prioritizing task accomplishment.

1 The research reported here was supported by NASA-Ames Research Center, Cooperative Agreement NCC-286 and by a contract with the Federal Aviation Administration, BAA80-005. Requests for reprints should be sent to Steven C. Predmore, NASA/University of Texas Crew Performance Project, 1609 Shoal Creek Blvd., Ste 200, Austin, TX 78701-1022.

2 The author wishes to thank Sean Maher for his assistance in transcribing and coding communications from the United 232 accident.

Action Decision Sequences. An important feature of our coding scheme is the capacity to classify verbal behavior in terms of *Action Decision Sequences (ADS).* Action Decision Sequences are linked communications centered on a single, situationally triggered event which requires *coordinated* action among crewmembers. The inclusion of the ADS code allows us to reduce crew interaction into a relatively limited number of behavioral sequences which effectively capture the multiple tasks faced by a crew. We were able to identify six ADSs which were common to both accidents:

Flight Control: Communications centered around efforts to maintain control of the aircraft; maintaining pitch attitude; use of differential engine thrust.

Damage Assessment: Communications focused on assessing the nature and extent of damage to the aircraft; identifying which systems are and are not operational.

Problem Solution: Communications dealing with corrective action; completing abnormal checklists; dumping fuel.

Landing: Discussions about potential landing sites; location of alternates; non-emergency landing preparations; manual extension of landing gear.

Emergency Preparations: Communications linked to preparing for an emergency landing; preparing the cabin; calling for equipment at the airport; reporting SOB's and fuel.

Social: Non-operational communications which address social-emotional and team-building concerns; introductions; tension release; affective support.

Distribution of Communication Across Tasks. Table 1 presents communications from both accidents broken down by *Action Decision Sequence*. For purposes of comparison, the distribution of communications for United 811 do not include interactions which occurred in the eight minutes prior to the loss of the cargo door. Not surprisingly, given the magnitude of structural damage suffered by both aircraft, the majority of communications in each case were concerned with Flight Control. The crew of United 811 was focused to a large degree on reducing their rate of descent to enable them to return to the Honolulu airport. For United 232, The differential engine thrust technique used by the crew to control the aircraft required highly precise and coordinated co-action which was for the most part a product of trial and error manipulations of the flight controls, and involved considerable verbal exchange.

ADS	United 811 (post-emergency)	United 232
Flight Control	35%	40%
Damage Assessment	20	15
Problem Solution	08	11
Landing	27	24
Emergency Prep.	10	07
Social	00	02

Table 1. Communications for United 811 and United 232 broken down by ADS.

Given the unique situation faced by the crew of UAL 232, the low percentage of communications devoted to Damage Assessment (15%) was initially surprising to us. However, it is possible that a good deal of damage assessment occurred in the 10 minutes immediately following the engine failure and was therefore not captured on the CVR. As a comparison, in the case of United 811, 80% of the communications directed toward Damage Assessment occurred within the first 10 minutes following the onset of the emergency. Finally, there was very little either crew could do in terms of prescribed corrective action and this is evidenced by the relatively low percentages of communication centered around Problem Solution.

We see 2% of communications in United 232 devoted to non-operational concerns. One of the unique aspects of the United 232 scenario is the active participation in the cockpit of the Check Airman who was initially riding as a passenger in the cabin. After volunteering his services he was asked by the Captain to assist in the cockpit, and in addition to overall statements of encouragement and affective support, some of the non-operational communications are a result of introductions between the Check Airman and the other crew members. Some might be critical of the presence of non-operational communication, such as interpersonal introductions, in the cockpit under high workload conditions. However, insofar as it does not interfere with operational concerns, such behavior can benefit performance through an enhanced sense of "team". Indeed, Ginnett (1987) found that one of the characteristics of effective Captains was a willingness to expand the boundaries of a crew to incorporate new members.

Distribution of tasks across crewmembers. As illustrated above, both crews were forced to contend with a number of tasks. This is especially difficult under conditions of high stress where an individual's perceptual focus tends to narrow, making the processing of multiple tasks particularly difficult. Under such conditions, teamwork is critical, and successful crew performance depends on the effective allocation of all available resources to the multiple task demands. We were interested, then, in how task focus, as measured through verbal behavior, was distributed among crew members. For purposes of brevity and because of the unique nature of a four-person cockpit, the remaining analyses will focus primarily on communications from the United 232 accident. Table 2 provides a breakdown of the communications of the crew of United 232 across the six ADSs.

ADS	CA	FO	FE	CK	OVERALL
Flight Control	43%	51%	16%	64%	40%
Damage Assessment	12	12	41	09	15
Problem Solution	09	06	29	10	11
Landing	26	22	09	14	24
Emergency Preparation	09	08	04	02	07
Social	02	01	00	01	02

Table 2. Percentage of each crewmembers' communication devoted to each Action Decision Sequence.

The decision by the Captain to invite the Check Airman to assist in the cockpit was cited by the NTSB as both "positive and appropriate" (NTSB/AAR-90/06, p. 76). The demands created by the use of differential engine thrust to control the aircraft made it very difficult for either the Captain or the First Officer to attend to other task demands. The presence of the Check Airman who was involved almost exclusively with Flight Control, freed the Captain to attend to other tasks when necessary. (The distribution of the Captain's communications across ADSs is explored in greater detail in the last section.) The table clearly shows the communications by the Check Airman (CK) were centered almost exclusively on Flight Control, whereas the Captain (CA) and First Officer (FO) split the bulk of their communications across the tasks of maintaining Flight Control and Landing. The Flight Engineer (FE) handled most of the radio communications with Dispatch and United Maintenance and this is reflected by the fact that 70% of his communications were concerned with Damage Assessment and Problem Solution. Perhaps the best evidence of the how this crew fully utilized available resources is provided in the first minutes following their awareness of the Check Airman's presence. In the two minutes after the Check Airman is invited into the cockpit, he is used by the crew in three critical ways: 1) he is immediately sent back into the cabin to do a visual inspection of exterior damage (Damage Assessment); 2) he is utilized for manipulation of the throttles (Flight Control); 3) he is asked for an update on the status of emergency preparations in the cabin (Emergency Prep.).

Task management over time. Our focus with this approach is not merely on the structure of communication with regard to multiple tasks, but also on the dynamics of multi-task processing. In addition to level and form of communication, the timing and interplay between the accomplishment of a number of tasks is critical to effective performance. Figure 3 illustrates the distribution of tasks (ADSs) over time by the crew of United 232 using an area chart format similar to Figures 1 and 2.

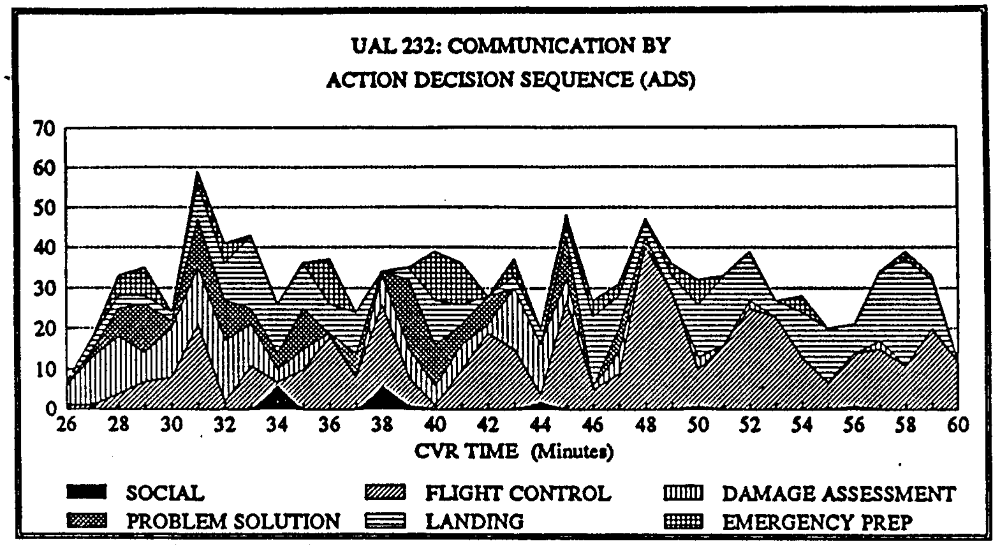


Figure 3. United 232: Breakdown of communications by ADS.

A close examination of this chart reveals two phases of task processing that emerge in this scenario. From Minute 26 to Minute 48 of the chart we see the communications of the crew consistently distributed across all five operational tasks. A notable feature of this period of the scenario is the peak of verbal activity that occurs at Minute 31. This represents the point where the Check Airman enters the cockpit after his visual damage inspection, and he is immediately brought into the loop with regard to damage to the flight control systems, corrective action that is ongoing, decisions about where to land, and instruction on the manipulation of the throttles.

After about Minute 48 however we see very little communication devoted to Damage Assessment or Problem Solution. At this point the aircraft is about 35 miles from the airport and the focus of the crew shifts from corrective action and assessing damage to landing the aircraft. The transition is marked by the following exchange between the Captain and the Flight Engineer:

47:23 Captain: What did SAM (United Airlines Maintenance) say, "good luck"?
47:24 Engineer: He hasn't said anything.
47:27 Captain: Okay. Well forget them. Tell 'em you're leavin' the air and you're gonna come back up here and help us . . .

This instruction by the Captain ensures that the focus of the crew is on the immediate and overriding concern at this point in the flight, landing the aircraft. It is obvious that there is little the crew can do in terms of corrective action, the crew is aware of which systems are operable, and there is no reason to expend additional crew resources on those tasks.

This chart also illustrates the distribution of non-operational communication throughout the scenario. The crew is informed of the presence of the Check Airman at Minute 29, and he is integrated into the flight control task at about Minute 31. There is a social exchange at Minute 34 which can be characterized as tension release and affective support. This exchange is completed with the Check Airman exclaiming, "We'll get this thing on the ground. Don't worry about it." Eleven seconds later the Captain made the decision to land at Sioux City. At Minute 38 the Captain, Check Airman and First Officer introduce one another. Throughout the scenario, non-operational communications occur during or immediately in the wake of a relatively low level of verbal activity which suggests that these exchanges did not jeopardize engagement in more critical operational activity.

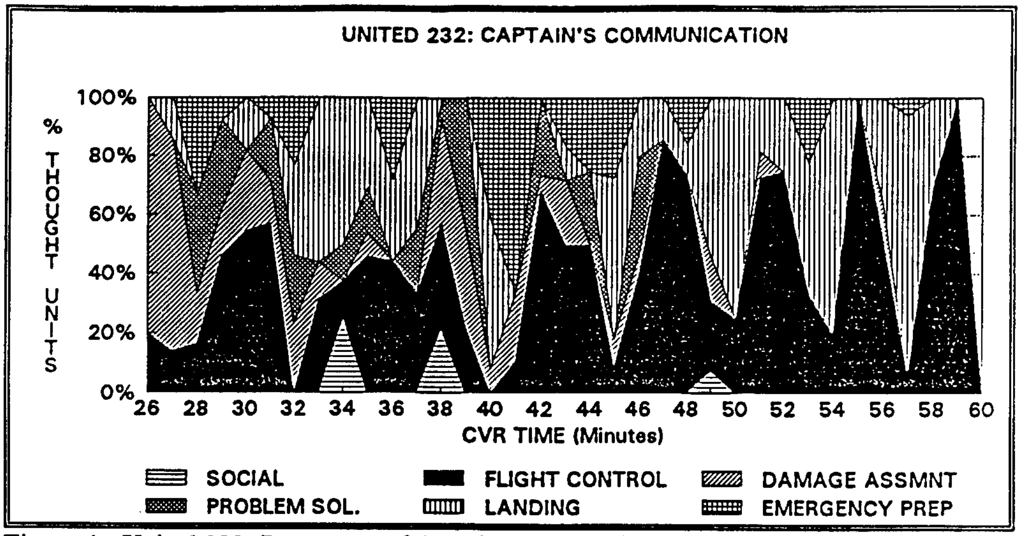


Figure 4. United 232: Percentage of Captain's communication by ADS over time.

Monitoring multiple task processing. As noted earlier, a primary benefit of the presence of the Check Airman in the cockpit was that this enabled the Captain to focus his attention not solely on matters related to Flight Control but to also allocate attention periodically to other ongoing concerns as well. Figure 4 is a 100% area chart which illustrates the percentage of Captain's communication devoted to each ADS over time. For example, at Minute 26 we see 20% of his verbal behavior was concerned with Flight Control and 80% was devoted to Damage Assessment. As time progresses, we see the Captain's communication increasingly devoted to Flight Control and Landing (i.e., after Minute 47). The most striking feature of this chart is the uniform pattern of peaks and valleys which indicate the regularity with which the Captain shifts the bulk of his communication from Flight Control issues (black areas) to other tasks.

SUMMARY

The analyses presented here clearly illustrate the impact of stressful events on the interactions of flight crews, as well as provide insights into the dynamics of effective crew coordination in the face of high workload, multi-task demands. The interactions of the crew of United 232 were marked by an efficient distribution of communications across multiple tasks and crewmembers, the maximum utilization of a fourth crewmember, the explicit prioritizing of task focus, and the active involvement of the Captain in all tasks throughout the scenario.

REFERENCES

Butler, R.E. (in press). Lessons from cross-fleet/cross airline observations: Evaluating the impact of CRM/LOS training. In R.S. Jensen (Ed.), Proceedings of the Sixth International Symposium on Aviation Psychology. Columbus, Ohio: Ohio State University.

Clothier, C.C. (in press). Behavioral interactions across various aircraft types: Results of systematic observation of line operations and simulations. In R.S. Jensen (Ed.), Proceedings of the Sixth International Symposium on Aviation Psychology. Columbus, Ohio: Ohio State University.

Ginnett, R.C. (1987). First encounters of the close kind: The formation process of airline flight crews. Unpublished Doctoral Dissertation. United States Air Force Academy.

National Transportation Safety Board. (1990a, April). Aircraft Accident Report (NTSB Report No. AAR-90-1). Washington, DC: NTSB Bureau of Accident Investigation. (NTIS No. PB90-910401).

National Transportation Safety Board. (1990b, November). Aircraft Accident Report (NTSB Report No. AAR-90-6). Washington, DC: NTSB Bureau of Accident Investigation. (NTIS No. PB90-910406).

Office of Aviation Medicine
Washington, D.C. 20591

Development of a Coding Form for Approach Control/ Pilot Voice Communications

O. Veronika Prinzo
Thomas W. Britton

Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma 73125

Alfred M. Hendrix

Contractor

May 1995

Final Report

U.S. Department of Transportation

Federal Aviation Administration

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-95/15		

4. Title and Subtitle	5. Report Date
Development of a coding form for approach control/pilot voice communications	May 1995
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
O.V. Prinzo, T.W. Britton, and A.M. Hendrix	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
FAA Civil Aeromedical Institute P.O. Box 25082 Oklahoma City, OK 73125	11. Contract or Grant No. DTFA-02-91-C-91089
12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aviation Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, D.C. 20591	14. Sponsoring Agency Code

15. Supplemental Notes

This work was performed under Task AM-D-94-HRR-117

16. Abstract

The Aviation Topics Speech Acts Taxonomy (ATSAT) is a tool for categorizing pilot/controller communications according to their purpose and for classifying communication errors. Air traffic controller communications that deviate from FAA Air Traffic Control Order 7110.65, and pilot communications that depart from the suggested communication in the Airman's Information Manual can be identified and labeled using the error codes provided in the ATSAT. By using the same procedures and tool to analyze communications, direct comparisons can be made between controller phraseology usage in the field and during simulation. Results of a preliminary study to measure inter-coder agreement revealed that novice coders were more dependent on the surface characteristics of the verbatim transcripts and experts relied more on domain specific background knowledge and experience with ATC phraseology to code ATC communications. If a researcher elects to use the ATSAT, we recommend that all coders receive the same orientation and instruction sessions prior to using the it.

17. Key Words	18. Distribution Statement
ATC-Pilot Communications Communication Taxonomy Phraseology	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
		34	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

DEVELOPMENT OF A CODING FORM FOR APPROACH CONTROL/PILOT VOICE COMMUNICATIONS

1.0 INTRODUCTION

1.1 Background

Voice-radio communication is central to air traffic control (ATC). Air traffic controllers are taught a standard phraseology[1] as part of their formal training, and once they are assigned to an air traffic control tower, terminal, or en route facility, their communication skills are reviewed periodically. Many government agencies, aviation industries, and researchers interested in controller/pilot communication often rely on the Aviation Safety Reporting System (ASRS) and the Office of Safety Information and Promotion (ASP) for aviation-related information. Verbal communication often is represented as a major category (with possibly several general types of communication topics) in addition to other controller performance measures on standardized FAA forms[2]. Voice-radio communication is included as part of investigations involving operational errors, system or pilot deviations, or other events that may have the potential to impact safety.

In aircraft-related accident investigations, a written verbatim transcript of the actual voice-radio communication is included as part of the official records to aid in the identification of the factors surrounding the incident. Written verbatim transcripts also are included in operational error/system deviation investigations. Some researchers (e.g., Cardosi, 1993; Morrow & Roldvold, 1994) have examined audio taped recordings of controller/pilot voice-radio communications provided by ATC. Transcribing and identifying potentially critical verbal communications can be an arduous and expensive task. A cost - effective approach is needed that would allow controller/pilot voice-radio communications to be coded and stored in a database for use by researchers and investigators to answer communication-based safety questions. In so doing, real progress could be achieved in understanding the dynamics of communication between controllers and pilots during routine operations and again when problems arise. A problem with existing databases is the lack of a uniform coding scheme which makes it difficult for users to gain a clear perspective of the magnitude of actual safety-related problems.

As part of a survey of the ATC/pilot voice communications literature, Prinzo and Britton (1993) included samples of air traffic control verbal communications taxonomies. Kanki and Foushee (1989) described typical flight crew performance and decision making (e.g., command, suggestion, inquiry, acknowledgment) using the speech act as the underlying unit of communication measure; whereas, Morrow, Lee, and Rodvold (in press) described TRACON[3] controller/pilot communication using the speech act and aviation topic (e.g., heading) in their analyses. A speech act is a single utterance used to convey a single action or intention for action (see glossary). In another approach Human Technologies, Inc. (1991) examined team co-ordination among en route controllers and pilots using the speech act to analyze communication patterns. Cardosi (1993) examined the complexity of en route communications by counting the number of elements (i.e., new pieces of information within a communication that increased memory load) in a transmission. Unfortunately, the results of these various efforts cannot be integrated and an overall conclusion reached since different measures were used.

From the Prinzo and Britton survey, it became apparent that different researchers used the same words to describe some communications; however,

[1] FAA Order 7110.65G Air Traffic Control
[2] FAA Form 7210-3, Final Operational Error/Deviation Report
[3] Terminal Radar Approach Control

the assigned meanings to those words were not always uniformly applied. For example, Golaszewski (1989) defined readback error as a loss in separation minima resulting from a controller's failure to detect (or correct) an incorrect readback by the pilot. Alternatively, Morrow, Lee, and Rodvold (1990) defined readback error as a failure to read back correctly the information contained in the original transmission; loss of separation was not considered. In some instances, words referencing concepts were provided without benefit of definition (e.g., frequency congestion) (Morrison & Wright, 1989) and left to the reader to interpret. It is uncertain whether experts and novices in the field of aviation consistently apply the same definitions to those words. Without benefit of uniform definitions, the risk of misunderstanding or misinterpretation increases.

1.2 Purpose

The purpose of the present research effort was to develop a voice communication taxonomy and method of data collection that could be used to analyze ATC/pilot voice-radio communication in a systematic and consistent fashion. That product is the Aviation Topics/Speech Acts Taxonomy (ATSAT). This taxonomy was developed as a tool for building a common ground of understanding of ATC communications through the use and application of a standard or common analytic procedure. The appropriateness of the ATSAT to other applications depends on the user's ultimate goal. Thus, the user will need to define the problem and determine the appropriate level of analysis. Within the ATSAT, the aviation topic presents a micro level of analysis and the speech act a macro level. In this taxonomy, the speech act defines the purpose of the utterance; that is, its intent.

The 5 speech act categories that make up the framework for the ATSAT and its corresponding coding form (See Appendix A) are: 1) Address, 2) Courtesy, 3) Instruction/-Clearance/Readback, 4) Advisory/Remark/Readback, and 5) Request/Readback. A sixth category, Non-Codable, is included as a general category. (See Appendix B.) Non-codable would include unintelligible transmissions due to equipment-related problems, delivery technique, and communications that could not be placed into any of the other major groupings.

The **aviation topic** is the basic unit of meaning (subject) and it is found within the speech act. Aviation topics place constraints on their associated speech acts by limiting the type of action that can occur. For example, headings, altitude restrictions, air speeds, and routes are aviation topics which are frequently included in transmissions containing instructions or requests. A complete list of aviation topics included in the ATSAT, along with their definitions, is included in this report. (See Appendix C.)

2.0 APPROACH

2.1 Development of the Aviation Topic Speech Act Taxonomy

A literature search was performed to acquire copies of the existing research conducted on controller/pilot voice radio communications. The speech act (Kanki & Foushee, 1989; Morrow, Clark, Lee, & Rodvold, 1990) was selected as the major type of communication element in a transmission under which the aviation topics were grouped. A list of the aviation topics was developed from the literature review for possible inclusion in the Aviation Topic Speech Act Taxonomy. These aviation topics were placed into the speech act category into which they were most likely to be found in a transmission.

Similarly, a list of the various types of communication problems was constructed from the Prinzo and Britton literature review and databases (e.g., ASRS). The communication problems were restricted to include only voice-radio messages between the controller and the pilot. Equipment related problems, such as faulty equipment, improperly worn headsets and microphones, intra-facility communication, inter-facility communication, and inter-flight-deck verbal communication were not included. Only controller/pilot voice radio communications within the terminal environment were addressed by this research.

Once the basic structure of the ATSAT was constructed, a sample of TRACON/pilot communications was obtained, transcribed, and coded using the taxonomy. Based on the VHF/UHF audio tapes provided, some of the speech acts were combined into a single category and several aviation topics were discarded or replaced. A retired controller served as the subject matter expert (SME) during the refinement of the ATSAT. FAA Order 7110.65G Air Traffic Control (1992), Airman's Information Manual (1992), and the FAA Order 7340.1M *Contractions* (1992) also were used as resources.

2.2 Identification of Problematic Verbal Communications

The Prinzo and Britton literature review aided in identifying message content errors and delivery technique errors as two major groups of communications-based problems. Although other types of communication problems have been identified (Morrison & Wright, 1989), many are equipment related problems (e.g., equipment outages, obsolete equipment). The ATSAT addresses only controller/pilot-centered verbal communication problems. Verbal communications, which deviated from standard phraseology specified in FAA Order 7110.65G or suggested pilot phraseology in the Airman's Information Manual, were grouped into those stemming from message content and delivery technique.

2.2.1 Message Content Errors

There are 7 different types of message content errors that are included on the ATSAT. These types of errors are listed in Table 1. Although grouped and seqential refer directly to numerical information, omission, substitution, and transposition, errors could also occur for other types of information, such as failing to include an aircraft callsign in a transmission where the callsign would be required. Substitution errors would include replacing the numbers in an assigned airspeed with the numbers assigned for a heading, or an altitude in a transmission that contained at least 2 aviation topics in a speech act instruction. Excessive verbiage errors include any words or phrases in addition to standard phraseology. Partial readbacks are similar to omission errors; however, partial readbacks occur when a pilot fails to include a piece of information in a readback. The two different codes are used because pilots and controllers are judged by the same phraseology standards for the ATSAT. According to FAA Order 7110.65G or the Airman's Information Manual, however, ATC phraseology is more rigidly prescribed for a controller than it is for a pilot.

2.2.2 Delivery Technique Errors

The analysis of the recorded voice-radio transmissions made by the master of the oil tanker Exxon Valdez served as a basis for defining delivery technique errors (Brenner & Cash, 1991). As displayed in Table 1, misarticulations (e.g., slurring of speech) and dysfluencies (e.g., hesitations) are the 2 major types of delivery technique errors included in the ATSAT. Misarticulations and dysfluencies have the potential for decreasing effective information transfer due to excessive pauses or the need to repeat a transmission.

3.0 PROCEDURE

3.1 Instructions

Table 2 lists the steps for transcribing, encoding, and entering the message content of audio transmissions onto the ATSAT Coding Form. Appendices A through D are provided to assist in the encoding process. Appendix A contains a copy of the coding form, a sample page of ATC/pilot transcribed communications, the same transcript page divided into aviation topics and coded with identified phraseology errors, and a completed copy of the coding form. Appendix B lists and defines each of the identified speech act categories according to their placement on the ATSAT Coding Form. Appendix C lists the aviation topics, along with their corresponding definition for each of the speech act categories, in the order of their occurrence on the ATSAT Coding Form. Appendices B and C should assist in the placement of message segments into their appropriate aviation topics and speech act categories on the ATSAT Coding Form. The definitions should not be confused with the more formal definitions of message content terms found in the glossary (Appendix E). Although there should be a close correspondence between how a message segment is defined and the category types presented on the ATSAT, the user occasionally may have to rely on personal experience when a message is slightly ambiguous. Appendix D lists some typical phraseology and delivery technique error types found in each aviation topic, along with their letter code; however, this is not an exhaustive list. It should also be noted that an aviation topic may contain more than one type of error.

Table 1

Communication Phraseology Errors in ATC/Pilot Transcripts.

Error	Code	Definition
Message Content Errors		
Grouped	G	Grouping of numerical information contrary to paragraph 2-85, FAA Order 7110.65G, March 1992
Sequential (Non-grouped)	N	Failure to group numbers in accordance with paragraphs 2-87, 2-88, 2-90, and non-use of the phonetic alphabet in accordance with paragraph 2-84, FAA Order 7110.65G, March 1992
Omission	O	Leaving out number(s), letter(s), word(s), prescribed in phraseology requirements in FAA Order 7110.65G, March 1992
Substitution	S	Use of word(s) or phrases(s) in lieu of phraseology outlined in FAA Order 7110.65G, March 1992 (e.g., "verify altitude" vs. "say altitude")
Transposition	T	Number(s) or word(s) used in the improper order (e.g., "TWA six forty-five" instead of "TWA five forty-six")
Excessive Verbiage	E	Adding word(s) or phrase(s) to phraseology outlined in FAA Order 7110.65G, March 1992, and the phraseology suggested in the Airman's Information Manual. (e.g., "TWA the number one airline six forty-five")
Partial Readback*	P	Pilot report or readback that does not include specific reference to a topic subject (i.e., altitude topic "out of six for four" would be recorded as a P. *Note: A verbatim readback of a controller's instruction or advisory would not be recorded as a P, nor would a readback containing a general acknowledgment and the aircraft identifier.
Delivery Technique Errors		
Dysfluency	D	Pause(s), stammer(s), utterance(s), that add no meaning to the message (e.g., "uh," "ah," or "ok" when not used as a general acknowledgment
Misarticulation	M	Improperly spoken words (i.e., slurs, stutters, mumbling, etc.)

For example, on line 11 of the "Sample Transcript Sheet" (cf. Appendix A, p. 6), air traffic control is transmitting the following message to Plato* 290:

"*Plato two-ninety roger clear visual three one left other traffic landing three one right.*"

The transcriptionist would spell out the numbers in the aircraft callsign and for each of the runways. Once transcribed, the message is segmented in each of the speech act categories by placing a diagonal slash between them (See Table 2, Part 2, Step 2): "*/ Plato two-ninety / roger / clear visual three one left / other traffic landing three one right /.*"

Next, each aviation topic in the transmission is numbered in the order in which it was spoken by the controller (See Table 2, Part 2, Step 3): "/ 1 *Plato two-ninety* / 2 *roger* / 3 *clear visual three one left* / 4 *other traffic landing three one right* /"

The final step in the encoding process is identifying those aviation topics containing errors (See Table 2, Part 2, Step 4). In the present example, an omission occurred in the third aviation topic, which should have read: "clear visual **approach runway** three one left," according to FAA Order 7110.65G. The error did not occur in the fourth aviation topic because that specific phraseology is not stated in the manual for issuing traffic advisories.

"/1 *Plato two-ninety* /2 *roger* /3 O *clear visual three one left* /4 *O; E other traffic landing three one right /*" should have been read as: "cleared visual approach runway three one left; traffic at (clock code, position, and aircraft type) landing runway three one right".

Once complete, the encoded message is transferred to the ATSAT Coding Form using the steps listed in Part 3 of Table 2. This is a fairly straight-forward process.

4.0 PRELIMINARY STUDY ON THE RELIABILITY OF THE ATSAT

4.1 Introduction

The ATSAT was developed by the authors to analyze phraseology usage by controllers and pilots at a micro level of analysis. It uses the terms and definitions found in FAA Order 7110.65 as its basic structure. The ATSAT may be helpful to other researchers in its current form or serve as a foundation or point of departure for developing their own voice communications coding schemes. To determine how reliable experts and novices were in coding ATC transmissions according to the ATSAT Coding Form's instructions and procedures, a preliminary study was performed.

4.2 Subjects

Four novices and 4 ATC instructors volunteered to code the same 25 transmissions from a transcript of ATC/pilot communications. Novices were FAA technical support staff who lacked domain specific prior knowledge of ATC terminology and phraseology usage. Experts were former ATCS employed as FAA Academy ATC instructors. Each volunteer was given a copy of the instructions from Tables 1 and 2 along with Appendices A through D to help with the coding.

4.3 Procedure

A 30 - minute orientation session on how to code the transmissions was given by one of the developers of the taxonomy who, as Facilitator, explained the coding process step by step with each group of novice and instructor coders. The novices were provided with 2 hours of additional instruction pertaining to ATC terminology and phraseology to ensure that they had the minimum requisite aviation knowledge necessary to complete the taxonomy. Since the Experts were responsible for observing and instructing their students on correct phraseology, they were not provided the additional instruction session.

4.4 Results and Discussion

The Facilitator also coded the same 25 transmissions to compare with the novices' and experts' data, and the percentage of items agreeing with the facilitator was computed. The coded transmissions of each group were compared to the coded transmissions of the facilitator for: (1) segmenting the entire message into speech acts and aviation topics, (2) correctly placing the segments onto the coding form, both in

* Plato was chosen for illustrative purposes only; It is a fictitious air carrier.

Table 2

Steps for Translating Audio-Taped Voice Communications to the ATSAT Coding Form.

1. Transcribe audio tapes to written verbatim copy.

Step 1. Identify and record the speaker identification.
Step 2. Copy message spelling out numbers.
Step 3. Enter time in minutes and seconds at the beginning of each transmission. (optional)
Step 4. Sequentially number transcript lines. (Each transmission should be numbered as a line. See example Appendix A.)

2. Encode transcript.

Step 1. Using Appendix C, divide each line of the transcript into aviation topics by placing a diagonal line at the beginning and end of each topic.
Step 2. Sequentially number the aviation topics, placing the number immediately after the beginning diagonal line.
Step 3. Using the "Communication Errors in ATC/pilot Transcripts Table" (Table 1), identify each error and place its letter code after its aviation topic number (Examples are provided in Appendix A.)

3. Transfer data to the ATSAT Coding Form. See Appendix A.

Step 1. Enter the facility name and the coder's name or initials in the appropriate spaces at the top of the ATSAT form.
Step 2. Record the line number from the transcript into the "Line No." column.
Step 3. Identify the speaker by entering the aircraft callsign for aircraft or "ATC" for the controller in the "Speaker" column of the ATSAT form.
Step 4. Sequentially number the communication attempts to a specific receiver and place that transmission number in the far right of the space in which the speaker is identified.
Step 5. Identify the receiver by entering the aircraft callsign for aircraft or "ATC" for controller in the "Receiver" column of the ATSAT form.
Step 6. Record each identified topic by entering the placement number of the topic transcript into the applicable topic column within the appropriate speech act category (Use the "Speech Act Categories" (Appendix B) and "Aviation Topics" (Appendix C) to determine the correct topics and categories.)
Step 7. Indicate any errors within the topics in the same space in which the topic is recorded, using the codes from the "Communication Errors in ATC/Pilot Transcripts" list (Appendix D).
Step 8. Place any additional information or explanation in the "Comment" column using the position number for reference.
Step 9. Repeat steps 8 & 9 until the entire line has been completed.
Step 10. Repeat steps 4 through 10 until each line from the transcript has been coded.

Table 3

Percentage Agreement by Novices and Experts with ATSAT Facilitator in Message Encoding and Classification.

		Message Encoding and Classification		
Coder	N	Segmentation	Placement into ATSAT Categories	Error Code
Experts	4	78%	30%	58%
Novices	4	89%	70%	73%

Table 4

Inter-rater Percentage Agreement in Placement of Message Segments into Speech Act and Aviation Topic Categories by Novices and Experts.

		ATSAT Category	
Coder	N	Speech Act	Aviation Topic
Experts	4	59%	56%
Novices	4	82%	78%

the proper speech act category and in the proper aviation topic, and (3) recognizing that a speech error occurred within an aviation topic. The coded transmissions of the novices then were compared to each other and percentage agreement was computed on properly placing the transmission segments into speech act categories and into aviation topics. The same comparison was performed for the experts.

As shown in Table 3, the novices and experts had higher percentage agreement on segmenting messages than they did on placing those segments into their respective categories on the ATSAT Coding Form or recognizing the presence of a speech error. Correct placement into ATSAT categories required that each segment be correctly labeled on the basis of speech act category and aviation topic and the correct placement of the coded information onto the coding form. It is not surprising that overall percent agreement decreased since a much more granular level of analysis is demanded here than on either segmentation or error recognition. Correct recognition of a speech error required the coders to simply compare the content of an aviation topic to the error type definitions and determine if a match occurred. On correctly recognizing a speech error within an aviation topic, the average agreement with the facilitator was higher for novices than for experts.

As shown in Table 4, novices had a higher percent agreement among themselves than the experts in placing transmission segments into the proper speech act and aviation topic categories. The differences between novices and instructors could have resulted from differences in ATSAT coding instructions. Instructors were not provided with the 2 hours of additional instruction pertaining to ATC terminology as were the novices. Novices could have approached the task from a similar perspective and purpose. The lack of formal instruction may have increased the variability among the instructors since they were forced to rely on their more subjective and individualized schemes for data classification. Also, they may have relied more on their prior knowledge and experience than on the materials provided to them; the former requiring less effort than the later.

5.0 DISCUSSION

The Aviation Topics-Speech Acts Taxonomy and coding form were developed for studying ATC/pilot voice communications. The ATSAT may be of use by other researchers in its present form or it may be modified to suit particular needs. If a researcher elects to use the ATSAT, several words of caution are in order that are not unique to the ATSAT. First, all coders should receive the same orientation and instruction sessions prior to using the ATSAT, regardless of their domain specific background knowledge or experience with ATC voice communications. Providing only the novices with the instructional session resulted in their being more in agreement with the Facilitator than were the instructors in labeling and placing the coded segments onto the coding form and identifying errors. Providing uniform orientation and instruction sessions to all coders should increase inter-coder agreement, since they would tend to approach the task from the same perspective and purpose.

Second, whereas the novices in the study were more dependent on the surface characteristics of the verbatim transcripts, the instructors may have relied more on experiential and domain specific knowledge to assist them in placing segments into their proper aviation topics and speech acts categories on the ATSAT coding form. Providing experts with instructions on the importance and use of objective measures over their subjective judgments when coding transmissions should improve inter-coder agreement.

Lastly, provisions for practice trials with direct feedback during training should increase inter-coder percentage agreement. The Facilitator was available while novices completed the ATSAT and provided further instruction upon request. Thus, immediacy of instruction, a common understanding of the concepts and procedures, and monitoring of performance may improve inter-coder percentage agreement.

6.0 REFERENCES

Airman's Information Manual, Official Guide to Basic Flight Information and ATC Procedures. (June, 1992). Washington DC: U.S. Government Printing Office.

FAA Order 7110.65G Air Traffic Control. (March, 1992). Washington DC: Air Traffic Rules and Procedures Service.

7340.1M *Contractions* (June, 1992). Washington DC: Air Traffic Rules and Procedures Service.

Brenner, M., & Cash, J. R. (1991). Speech analysis as an index of alcohol intoxication - The Exxon Valdez accident. *Aviation, Space, and Environmental Medicine,* **62**, pp. 893-898.

Cardosi, K., M. (1993). An Analysis of En Route Controller-Pilot Voice Communications. Springfield, VA: National Technical Information Service.

Golaszewski, R. (1989). An analysis of pilot-controller-read-back errors. *Journal of ATC,* (October-December), 53-56.

Human Technology, Inc. (1991). Analysis of controller communication in en-route air traffic control. Report to the Federal Aviation Administration. McLean, VA.

Kanki, B.G., & Foushee, H.C. (1989). Communication as group process mediator of aircrew performance. *Aviation, Space, and Environmental Medicine,* **60**, pp. 402-410.

Morrison, R. & Wright, R. H. (April, 1989). ATC Control and communications problems: An overview of recent ASRS data. In (Ed: R. S. Jensen) *Proceedings of the Fifth International Symposium of Aviation Psychology,* Columbus, Ohio; Volume 2.

Morrow, D., Clark, H., Lee, A., & Rodvold, M. (Dec., 1990). Collaboration in Controller-Pilot Communications. Presented at the 1990 Meeting of the Psychonomics Society, New Orleans, LA.

Morrow, D., Lee, A., & Rodvold, M. (Dec., 1990). Analysis of Routine Pilot-Controller Communications. In *Managing the Modern Cockpit: Third Human Error Avoidance Techniques Conference Proceedings.* Warrendale PA: Society of Automotive Engineers, Inc.

Morrow, D., Lee, A., & Rodvold, M. (in press). Analysis of problems in routine controller-pilot communications. *International Journal of Aviation Psychology.*

Prinzo, O. V. & Britton, T. W. (1993). *ATC/Pilot Voice Communications - A Survey of the Literature.* Washington, DC: Office of Aviation Medicine Report no. DOT/FAA/AM-93/20.

Sample Transcript Sheet

1	00:00	PLATO 754	ZERO TWO ZERO PLATO SEVEN FIVE FOUR
2	00:03	ATC	PLATO THIRTY-FIVE HEAVY CONTACT (NAME) TOWER ONE TWO THREE POINT FOUR GOOD DAY
3	00:09	PLATO 35	PLATO THIRTY-FIVE GOOD DAY AND THANK YOU A LOT
4	00:24	ATC	PLATO SEVEN FIFTY-FOUR SAY YOUR SPEED
5	00:32	PLATO 754	AH WE'RE DOING ONE NINETY SEVEN FIFTY-FOUR
6	00:38	ATC	SEVEN FIFTY-FOUR ROGER INCREASE SPEED TO TWO ONE ZERO
7	00:41	PLATO 754	PICK IT UP TO TWO TEN SEVEN FIFTY-FOUR
8	00:47	PLATO 290	APPROACH PLATO TWO-NINETY AT A FOUR POINT SIX FOR TWO
9	00:48	ATC	PLATO TWO-NINETY (NAME) APPROACH TURN LEFT HEADING ZERO TWO ZERO
10	00:52	PLATO 290	ZERO TWO ZERO WE HAVE THE AIRPORT IN SIGHT ALSO
11	00:56	ATC	PLATO TWO-NINETY ROGER CLEAR VISUAL THREE ONE LEFT OTHER TRAFFIC LANDING THREE ONE RIGHT
12	00:59	PLATO 290	CLEAR TO VISUAL THREE ONE LEFT AND WE'LL WATCH THE TRAFFIC ON THE RIGHT ONE PLATO TWO-NINETY
13	01:04	ATC	ATTENTION ALL AIRCRAFT LANDING (NAME) INFORMATION PAPA NOW CURRENT THE WEATHER IS STILL BETTER THAN FIVE THOUSAND FIVE
14	01:24	PLATO 880	(TRANSMISSION PARTIALLY BLOCKED) SIX THOUSAND SEVEN HUNDRED FOR THREE THOUSAND HEADING ZERO FOUR ZERO

Facility:______________________ Coder:____________________

			Address/Addressee		Courtesy			Instruction/Clearance - Readback/Acknowledgment				
			Speaker	Receiver					Heading		Altitude	
Line No.	Speaker	Receiver	ID	ID	Thanks	Greetings	Apology	Heading	Mod.	Altitude	Restrict	Speed

ATSAT Coding Form

Facility:________________________ Coder:____________________

Instruction/Clearance - Readback/Acknowledgment (con't)						Advisory/Remarks - Readback/Acknowledgment						
Appr/ Departure	Freq.	Holding	Route	Transpond Code	General Acknowl	Heading	Heading Mod.	Altitude	Altitude Restrict	Speed	Appr/ Departure	Route Position

ATSAT Coding Form

Facility:____________________ Coder:____________________

Advisory/Remarks - Readback/Acknowledgment (con't)						Request - Readback/Acknowledgment						
			General	Traffic	General				Appr/	Route		
NOTAMS	ATIS	Weather	Sighting	Info	Acknowl	Heading	Altitude	Speed	Departure	Position	Type	NOTAM

ATSAT Coding Form

Facility:_______________ Coder:_______________

Request - Readback (con't)			Non - codable			
		General				
Traffic	Weather	Acknowl	Equipment	Delivery	Other	Comments

ATSAT Coding Form

Sample Transcript Sheet

1	00:00	PLATO 754	/ {1P} ZERO TWO ZERO / {2N} PLATO SEVEN FIVE FOUR /
2	00:03	ATC	/{1} PLATO THIRTY-FIVE HEAVY / {2}CONTACT (NAME) TOWER ONE TWO THREE POINT FOUR / {3} GOOD DAY /
3	00:09	PLATO 35	/ {1} PLATO THIRTY-FIVE / {2} GOOD DAY / {3E} AND THANK YOU A LOT /
4	00:24	ATC	/ {1} PLATO SEVEN FIFTY-FOUR / {2E} SAY YOUR SPEED /
5	00:32	PLATO 754	/ {1DC} AH WE'RE DOING ONE NINETY / {2P} SEVEN FIFTY-FOUR /
6	00:38	ATC	/ {1O} SEVEN FIFTY-FOUR / {2} ROGER / {3O} INCREASE SPEED TO TWO ONE ZERO /
7	00:41	PLATO 754	/ {1SCP} PICK IT UP TO TWO TEN / {2P} SEVEN FIFTY-FOUR /
8	00:47	PLATO 290	/ {1P} APPROACH /{2} PLATO TWO-NINETY / {3EP} AT A FOUR POINT SIX FOR TWO /
9	00:48	ATC	/ {1} PLATO TWO-NINETY / {2} (NAME) APPROACH / {3} TURN LEFT HEADING ZERO TWO ZERO /
10	00:52	PLATO 290	/ {1P} ZERO TWO ZERO / {2E} WE HAVE THE AIRPORT IN SIGHT ALSO /
11	00:56	ATC	/ {1} PLATO TWO-NINETY / {2} ROGER / {3O} CLEAR VISUAL THREE ONE LEFT / {4OE} OTHER TRAFFIC LANDING THREE ONE RIGHT /
12	00:59	PLATO 290	/ {1} CLEAR TO VISUAL THREE ONE LEFT / {2S} AND WE'LL WATCH THE TRAFFIC ON THE RIGHT ONE / {3} PLATO TWO-NINETY /
13	01:04	ATC	/ {1} ATTENTION ALL AIRCRAFT LANDING (NAME) / {2} INFORMATION PAPA NOW CURRENT / {3OE} THE WEATHER IS STILL BETTER THAN FIVE THOUSAND FIVE /
14	01:24	PLATO 880	/ {1} (TRANSMISSION PARTIALLY BLOCKED) / {2} SIX THOUSAND SEVEN HUNDRED FOR THREE THOUSAND / {3} HEADING ZERO FOUR ZERO /

Facility: ____________________ Coder: ____________________

			Address/Addressee		Courtesy			Instruction/Clearance - Readback/Acknowledgment				
			Speaker	Receiver					Heading		Altitude	
Line No.	Speaker	Receiver	ID	ID	Thanks	Greetings	Apology	Heading	Mod.	Altitude	Restrict	Speed
1	PLATO 754	ATC	2N					1P				
2	ATC	PLATO 754		1		3						
3	PLATO 35	ATC	1		3E	2						
4	ATC	PLATO 754		1								
5	PLATO 754	ATC	2P									
6	ATC	PLATO 754		1O								3O
7	PLATO 754	ATC	2P									1SCP
8	PLATO 290	ATC	2	1P								
9	ATC	PLATO 290	2	1				3				
10	PLATO 290	ATC	1P									
11	ATC	PLATO 290		1								
12	PLATO 290	ATC	3									
13	ATC	ALL		1								
14	PLATO 880	ATC										

ATSAT Coding Form

Facility:_______________ Coder:_______________

Instruction/Clearance - Readback/Acknowledgment (con't)						Advisory/Remarks - Readback/Acknowledgment						
Appr/ Departure	Freq.	Holding	Route	Transpond Code	General Acknowl	Heading	Heading Mod.	Altitude	Altitude Restrict	Speed	Appr/ Departure	Route Position
	2											
								3EP				
3O												
1												
						3		2				

ATSAT Coding Form

Facility:____________________ Coder:____________________

Advisory/Remarks - Readback/Acknowledgment (con't)						Request - Readback/Acknowledgment						
			General	Traffic	General				Appr/	Route		
ATIS	NOTAMS	Weather	Sighting	Info	Acknowl	Heading	Altitude	Speed	Departure	Position	Type	NOTAM
								2E				
								1DC				
			2E									
				4OE	2							
				2								
2		3OE										

ATSAT Coding Form

Facility:________________________ Coder:____________________

Request - Readback (con't)			Non - codable			
		General				
Traffic	Weather	Acknowl	Equipment	Delivery	Other	Comments
		2				
			1			Transmission partially blocked

ATSAT Coding Form

APPENDIX B: DEFINITION OF SPEECH ACT CATEGORIES IDENTIFIED IN 10 HOURS OF ATC/PILOT TRANSCRIPTS

1. **Address/Addressee.**

   The facility/position or aircraft identified as speaker or receiver (e.g., (Facility Name) TRACON, (Facility Name) departure, sector twenty-one, Plato two forty-one, November one two three alpha, Baron one two three alpha).

2. **Courtesy.**

   Word(s) or phrase(s) spoken as an act of courtesy.

3. **Instruction/Clearance—Readback/Acknowledgment.**

   Instruction/Clearance: Phraseology used by a controller to issue instructions to an aircraft (e.g., climb and maintain three thousand, turn left heading two two zero, cleared ILS runway three five right approach).

   Readback/Acknowledgment: Words or phrases spoken by a pilot or controller ***in response to*** an instruction/clearance.

4. **Advisory/Remark—Readback/Acknowledgment.**

   Advisory/Remark: Required communication based on the controller's responsibility for issuing advisories (e.g., altimeter, traffic, expected approach or altitude, a request for information, etc.) and the pilot's responsibility for making certain reports (e.g., ATIS, position, altitude, speed, etc.).

   Readback/Acknowledgment: Words or phrases spoken by a pilot or controller ***in response to*** an advisory/ remark.

5. **Request—Readback/Acknowledgment.**

   Request: Speech act initiated by the pilot or controller for the purpose of acquiring information and/or a service.

   Readback/Acknowledgment: Words or phrases spoken by a pilot or controller ***in response to*** a request.

6. **Non-Codable Remarks.**

   Remarks/comments that are not codable into a speech act of **Address/Addressee, Courtesy, Instruction/ Clearance Readback/Acknowledgment, Advisory/Remark—Readback/ Acknowledgment, Request—Readback/Acknowledgment**. A speech act that is unintelligible due to equipment problems or speaker delivery.

7. **Comments.**

   Information entered by encoder to clarify a coding entry.

APPENDIX C: AVIATION TOPICS WITHIN THE SPEECH ACT CATEGORIES

1. **Address/Addressee.**

a. **Speaker:** Identification of the speaker.

b. **Receiver:** Identification of the receiver.

2. **Courtesy.**

a. **Thanks:** "Thanks," "thank you," or words of appreciation.

b. **Greetings:** "Good day," "so long," "hello".

c. **Apology:** Any apology, example: "I'm sorry," "I owe you," etc.

3. **Instruction/Clearance—Readback/Acknowledgment.**

a. **Heading:** An assigned vector or readback by a pilot.

b. **Heading Modifier:** A word or phrase indicating an increased/decreased rate of turn.

c. **Altitude:** Altitude assigned by a controller or readback by a pilot.

d. **Altitude Restriction:** Any restriction to altitude assignment by a controller or readback by a pilot. *Note:* Includes "no delay in descent".

e. **Speed:** Speed assigned by a controller or readback by a pilot. *Note:* "Present speed," "reduce now," are speed assignments.

f. **Approach/Departure:** A clearance given by a controller to make an approach to an airport, or runway assignment (either IFR or VFR) or readback by a pilot.

g. **Frequency:** A radio frequency used for communications or navigation aid assignment by a controller or readback by a pilot. *Note:* May or may not include megahertz frequency.

h. **Holding:** Holding instruction issued by a controller or readback by a pilot.

i. **Route:** Any instruction issued by a controller that pertains to the course an aircraft is assigned or readback by a pilot. *Note:* Includes headings, vectors, airways, J routes, ILS, approaches, departure and arrival routes (SID, STAR, PDR).

j. **Transponder:** A beacon code and/or ident instructions issued by a controller or readback by a pilot.

k. **General Acknowledgment:** Word(s) used by a pilot as general acknowledgment of a clearance/ instruction. *Note:* "Roger," "ok," "alright," may be used in addition to aircraft identification and/or readback of all or portions of a clearance/instruction.

4. **Advisory/Remark—Readback/Acknowledgment.**

a. **Heading:** An expected vector/heading given by a controller or his/her readback of a pilot report. A pilot report of a vector/heading.

b. **Heading Modifier:** Word(s) or phrase(s) used by either a controller or pilot indicating an increased/ decreased rate of turn.

c. **Altitude:** An expected altitude assignment issued by a controller or his/her acknowledgment of an altitude reported by a pilot. An altitude reported by a pilot.

d. **Altitude Restriction:** An expected altitude restriction issued by a controller or his/her readback of a report by a pilot. A pilot report of an altitude restriction.

e. **Speed:** An expected speed assignment issued by a controller or his/her readback of a pilot speed report. A speed reported by a pilot.

f. **Approach/Departure:** An expected approach/ departure instruction issued by a controller or his/her readback of a pilot report. A pilot report of assigned approach/departure.

g. **Route/Position:** A route or position issued by a controller or his/her readback of a route or position reported by a pilot. A pilot report of a route or position.

h. **NOTAM/Advisory:** A Notice to Airmen (NOTAM) or aviation advisories issued by a controller or his/her readback of a pilot report. A pilot report of aviation advisories or his/her readback of a NOTAM/advisory (e.g., runway construction, status of navigation equipment, bird traffic.).

APPENDIX D:
SOME TYPICAL ERRORS WITHIN SPEECH ACT TOPICS

A. **Speaker:** Reference par 2-76, 77, 86, 87 of FAA Order 7110.65G and par 4-33 of AIM
Example - Initial contact:
Pilot: "Regional Approach **Plato ten twenty-two...**"
Controller: "Plato ten twenty-two **Regional Approach...**"

Example - After initial contact:
Pilot: "**Plato ten twenty-two...**"
Controller: "Plato ten twenty-two..."

		CODE
1.	Omission of facility name or function	O
2.	Omission of company name, general aviation designator, military service, etc.	O
3.	Omission of any number in the identification or use of less than three numbers/letters in general aviation or military identification	O
4.	Failure to group air carrier callsigns or to use the phonetic alphabet in aircraft identifications	N
5.	Grouping military or general aviation callsigns	G
6.	Additions to callsigns	E
7.	Substitution of company name, military service, or complete numbers/letters, etc.	S
8.	Transposed numbers/letters	T

B. **Receiver:** Reference par 2-76, 77, 86, 87 of 7110,65G and 4-33 of AIM
Example - Initial contact:
Pilot: "**Regional Approach** Plato ten twenty-two..."
Controller: "**Plato ten twenty-two** Regional Approach..."

Example - After initial contact:
Controller: "**Plato ten twenty-two...**"
Pilot: Ground station (control facility) may be omitted

		CODE
1.	Omission of facility name or function	O
2.	Omission of company name, general aviation designator, military service, etc.	O
3.	Omission of any number in identification or use of less than three numbers/letters in general aviation or military identification	O
4.	Failure to group air carrier callsigns or to use the phonetic alphabet in aircraft identifications	N
5.	Grouping military or general aviation callsigns	G
6.	Additions to callsign	E
7.	Substitution of company name, military service, or complete numbers/letters, etc.	S
8.	Transposed numbers/letters	T

Note: A pilot readback of controller's exact instructions is not recorded as an error.

CODE

1. Word(s) in lieu of "expedite" or "immediately" .... S
2. Failure to identify runway or NAVAID
   by the controller .... O
   by the pilot .... P
3. Errors may include those listed in E. Altitude.

G. **Speed:** Reference par 2-851, 5-101 of FAA Order 7110.65G and 4-41, 86, 91 of AIM
Example:
   Controller: "**...maintain present speed**"
   Pilot: "**...(number of knots) knots**"

CODE

1. Omission of "knots," except when assigning a speed in conjunction with an altitude .... O
2. Omission of "knots" or "speed" by pilot .... P
3. Grouping of speed numbers .... G
4. Additional and unnecessary words .... E

Note: One method of speed control not obvious, but used at least twice, was the assignment of altitude to allow higher speed or force a lower speed.

H. **Approach/Departure:** Reference par 2-85j, 4-60, 80, Chapter 5 sections 9-10, par 7-2, 10, par 7-2, 10, 31, 32, 33, 111 of FAA Order 7110.65G and par 4-86 of AIM.
Example:
   Controller: "**...cleared ILS runway three five left**"
   Pilot: "**...ILS runway three five left approach**"

CODE

1. Grouping of runway numbers .... G
2. Incomplete description of approach by controllers .... O
3. Incomplete description of approach by pilot .... P
4. Use of "join" for "intercept" and vice versa .... S

I. **Frequency:** Reference par 2-85k, 86 of FAA Order 7110.65G and 4-33d of AIM
Example:
   Controller: "**...contact (Facility) tower one one eight point five**"
   Pilot: "**...(Facility) tower one one eight point five**"

CODE

1. Addition of "on," "now," "the," etc. .... E
2. Grouping of frequency numbers .... G
3. Omission of "point"
   by the controller .... O
   by the pilot .... P
4. Omission of the facility name or function by the controller .... O

P. **Weather:** Reference par 2-111, 2-85f, RVR 2-122.

		CODE
1.	Omission of "runway" when giving RVR	O
2.	Grouping numbers contrary to standard phraseology	G
3.	Non-grouping of numbers contrary to standard phraseology	N
4.	Failure to include the station (altimeter or weather)	O

Q. **ATIS:** The pilot should report his awareness of current airport information (ATIS) by stating the phonetic letter of the ATIS information he has received. Controller communication reference to ATIS should be to confirm pilot awareness. Specific phraseology is not provided in either AIM or FAA Order 7110.65G.

		CODE
1.	Addition to a single phonetic letter	E
2.	Non-phonetic or incorrect phonetic letters	S
3.	Words/phrases other than "confirm ATIS (letter)	S

R. **General Acknowledgment:** Word(s) used by a pilot as a general acknowledgment of a clearance/ instruction.
Note: "Roger," "ok," "alright," may be used in addition to aircraft identification and/or readback of all or portions of a clearance/instruction.

Note: This appendix is added as a guide for coding communication errors onto the ATSAT Coding Form. The lists of errors are not exhaustive, and it is possible to have more than one error per aviation topic. Controller standard phraseology is taken from applicable parts of FAA Order 7110.65G, dated March 5, 1992. Pilot phraseology is taken from applicable parts of AIM, dated March 5, 1992, and where no phraseology is listed, a combination of FAA Order 7110.65G and par 4-86b1 and 4-86b2 of the AIM is used. The examples are illustrations of correct phraseology, and the underlined portions refer to the aviation topics. Aviation topics appear in bold type.

NASA Contractor Report 177634

Activity Catalog Tool (A.C.T.) v2.0 User Manual

Leon D. Segal and Anthony D. Andre

A. C. T.
Activity Catalog Tool

Version 2.0ß
(January '94)

- A.C.T. was originally designed and written by: Leon D. Segal
(A.C.T. version 1.0 © 1991, Leon D. Segal)

- Version 2.0 was designed by: Leon Segal and Anthony Andre
written by: Dominic Wong

- When writing papers in which A.C.T. is used or mentioned, please cite:
Segal, L.D., and Andre, A.D. (1993). *Activity Catalog Tool (A.C.T.) v2.0 User Manual.* NASA Contractor Report CR-177634. Moffett Field, CA: NASA ARC.

- For updates, questions and to report bugs, please contact:
Leon Segal or Anthony Andre, NASA ARC, MS 262-3, Moffett Field, CA 94035.
e-mail: leons@eos.arc.nasa.gov OR andre@eos.arc.nasa.gov

- A.C.T. version 2.0 was developed with the support of:
NASA Ames Research Center, Aerospace Human Factors Research Division, FLR Branch; Sterling Software and Western Aerospace Laboratories, Inc.

CONTRACT NCC2-486
January 1994

NASA
National Aeronautics and
Space Administration

NASA Contractor Report 177634

Activity Catalog Tool (A.C.T.) v2.0 User Manual

Leon D. Segal and Anthony D. Andre

Western Aerospace Labs, Inc.
16111 Mays Avenue
Monte Sereno, CA 95030

Prepared for
Ames Research Center
CONTRACT NCC2-486
January 1994

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035-1000

A-78

User Documentation for Activity Catalog Tool (A.C.T.) v2.0ß

A.C.T. is a tool for recording and analyzing sequences of activity over time. It was designed as an aid for professionals who are interested in observing and understanding human behavior in field settings, or for the study of video or audio recordings of the same. Specifically, the program is aimed at two primary areas of interest: human-machine interactions and interactions between humans. The program provides a means by which an observer can record an observed sequence of events, logging such parameters as frequency and duration of particular events. The program goes further by providing the user with a quantified description of the observed sequence through application of a basic set of statistical routines. Finally, the program enables merging and appending of several files and more extensive analysis of the resultant data.

In order to best explain the utility and potential of A.C.T., we have programmed a demonstration file which is included on the A.C.T. disk. This file, along with the following set of instructions and procedures, will serve as your introduction to A.C.T. We encourage you to open the demonstration file ("DEMO.A.C.T."), and follow the step-by-step tutorial provided below.

About A.C.T.

• The version of A.C.T you have received (v2.0ß) is the first public release of the software. While much effort has been spent eliminating any bugs, we acknowledge that, as with any new software, we can not guarantee bug-free operation. Accordingly, keep in mind that we depend on you for feedback concerning any problems with, or questions about, this software. For updates, questions, and to report bugs, please contact Leon Segal or Anthony Andre, NASA ARC, MS 262-3, Moffett Field, CA 94035. You may also reach us via E-mail: leons@eos.arc.nasa.gov OR andre@eos.arc.nasa.gov

- When writing articles or reports in which A.C.T. is used or mentioned, please cite:
  Segal, L.D. and Andre, A.D. (1993). *Activity Catalog Tool (A.C.T.) v2.0 User Manual.* NASA Contractor Report CR-177634. Moffett Field, CA: NASA Ames Research Center.

- We would appreciate receiving a copy or citation of any articles or reports in which A.C.T. is referenced.

A.C.T. Program Requirements

- Mac II class (68020) or higher
- System 7.0 or higher (if you want post-processing "drag-and-drop" capability)
- Working copy of Microsoft® Word© or any text processor for viewing data files

DEMO.A.C.T.

Since A.C.T. is designed to be used as a tool during field observations and video analysis, we have picked a particular scenario - the office - to serve as the example for the illustration of the program's functions. For the purpose of this demonstration, imagine that we were interested in recording and analyzing the activities of a person in their office: we may be interested in designing a new layout for the office, or providing office personnel with a new type of information technology. We have selected eight categories of behavior which most interest us: the person's physical position (standing or sitting), seven tasks in which they may engage (writing, typing, reading paper documents, reading the computer screen, searching through files or talking on the phone), and one event which may be important to note - a visitor entering the office.

The following description of our operation of A.C.T. assumes that we are sitting in an office, or watching a closed-circuit TV or pre-recorded video of the same, observing an individual interact with the physical environment which comprises that office.

Running DEMO.A.C.T.

Important note for Powerbook users: Your Powerbook has several settings which help it conserve power; these same settings will cause the graphics used in A.C.T. to look as if the clock is not running smoothly. Note that this effect is visual only, and does not, in fact, effect the program's clock in any way. For the sake of viewing a smooth visual interface, however, you may want to follow these steps:

1. *From the Apple menu, select: Control Panels*
2. *From the Control Panels menu, select: PowerBook*
3. *Hold down the option key and click on the "Options..." button in the Battery Conservation box*
4. *a. Select: "Don't sleep when plugged in" (if you intend to use external power)*
   b. Select: "Don't allow cycling"
   c. Select: "Standard speed"

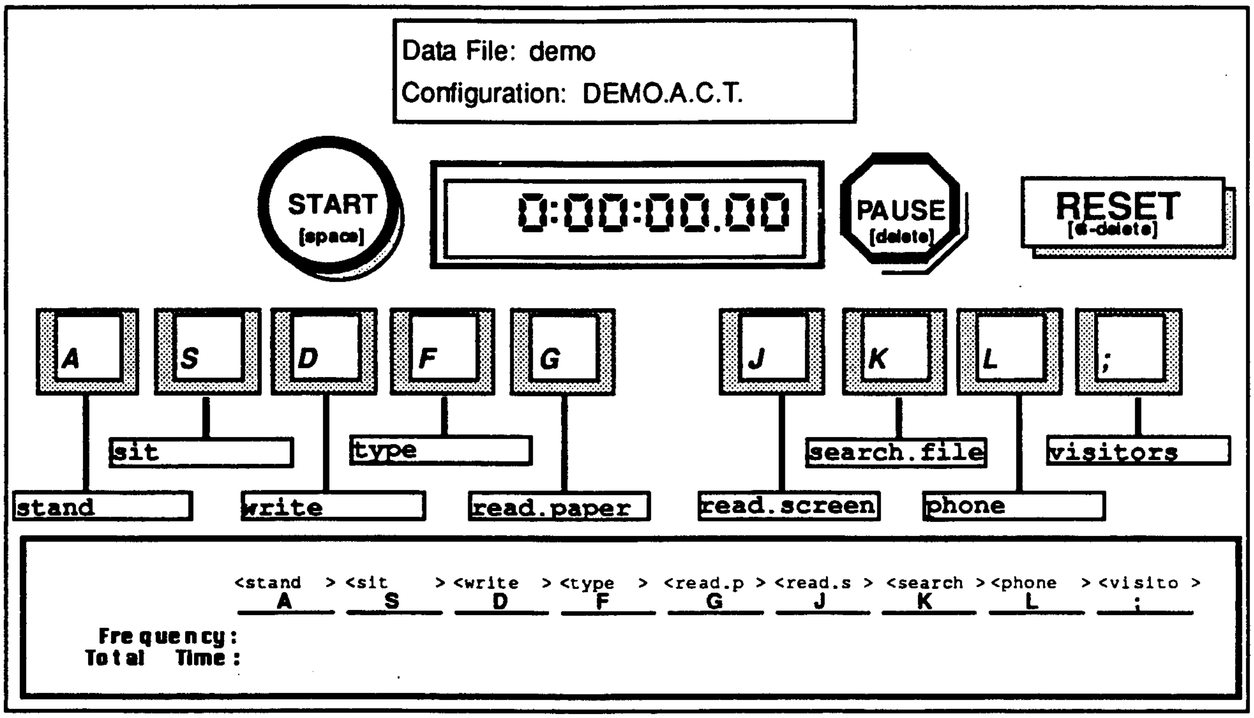
To start the demo program (DEMO.A.C.T.)

- Double click on the "DEMO.A.C.T." file icon: this opens our previously-defined office configuration.

- Press OK or hit the return key to pass the title screen

- Enter "demo" for the data file and select Save or hit the return key: you have now named the file in which the next session's data will be collected.
  Note: You do not have to enter a file name to go to the next screen (by selecting the Cancel button), but you will not be able to start running the session until a data file is named.

- You are now looking at A.C.T.'s data-collection interface:



Notice that the keyboard-like interface is configured for those behaviors and events on which the demonstration study focuses. Each one of the nine activities (measurements, or observational categories) described above is assigned to one of the nine keys on the display. On the screen, each key is attached to a label which describes the particular activity (measure) to which the key is mapped.

The *Statistics Box* appears in the bottom part of the configuration screen, displaying nine columns (corresponding to the nine configuration keys), providing the two fundamental counts of Frequency and Total Time for each.

You are now ready to start the observation session.

Start:
[space bar]

To start the observation session: Press the space bar or use the cursor to select the "Start" button. The clock starts running and all data-collection keys are activated. While the clock runs, every key press will be entered as a separate line in the data file.

Key functions:
Different keys may serve different goals.

Collecting data: Let's assume that the observed person is sitting at their desk when the session begins. Press the "S" key — which is labeled "sit" — and notice the feedback: a click sounds with the key press; a black tab appears on the key and will remain there as long as it is selected; information in the statistics box at the bottom of the screen indicates that the "sitting" button has been pressed once, as well as continuously updating the length of time that button has been selected.

Notice that since the keys are configured to resemble the nine keys in the "home" position on a keyboard, the interface is designed to afford "blind" dedication of your fingers to the keys, thus allowing you to enter data without looking at the keyboard.

In this demonstration, the "S" key was configured to measure both frequency and duration of sitting. Depending on the research questions and scenario, keys are configured to perform particular functions. Different keys may serve different goals, as you will see in the following section.

Serial keys:
These keys record events and behaviors that are mutually exclusive.

Now the observed person stands up. Press the "A," or "stand," button. Notice that along with the click sound, the black tab on the key, and the information in the statistics table, one more thing has occurred: the "sit" (S) key has been switched off. From a theoretical point of view, this is obvious — the observed person cannot be seated and standing at the same time.

In the A.C.T. language, the two keys of "sit" and "stand" are considered *serial keys*. Serial keys are used to catalog behaviors and events that are mutually exclusive — only one can occur at any given time. As you will see later, any combination of keys — from two to nine keys — can be configured as serial. There can be more than one grouping of serial keys as well. For example, in this office configuration, you will notice that the read.paper (G) and read.screen (J) keys are also configured as serial.

Imagine the person standing and sitting several times and record those activities by alternating between the A and S keys. Notice the changes in frequency and time measurements displayed in the statistics box at the bottom of the screen.

Parallel keys: *These keys record events and behaviors that may occur at the same time.*

Let's assume that our subject has started talking on the phone. Press the "L" key — labeled "phone" — to create a record of our observation of the person talking. Notice that one press of the key turns it on; it will remain so until you press it again when they have finished talking. As long as the key is "on" you will see the black tab, as well as the incrementing of time in the corresponding cell in the statistics box (row: Total Time On / column: J).

Now let's assume that the person starts to search through the file drawer while still on the phone. Press the "K" key — labeled "search.file" — to create a record of our observation of the person searching for a file.

Notice that pressing the K key did not affect the status of the previously-selected "phone" (L) key. This correctly reflects the fact that one can talk on the phone and search for a file at the same time. Accordingly, in the A.C.T. language, the "phone" and "search.file" keys are considered *parallel keys*. <u>Parallel keys are used to catalog behaviors and events that can take place simultaneously.</u>

Notice also that pressing either the "phone" or "search.file" key did not affect the status of the previously-selected "sit" (S) key. This correctly reflects the fact that talking on the phone and searching the file drawer did not alter the fact that the person is still sitting. The person may talk on the phone while sitting or while standing; they may type while reading from paper (when copying from a book) or while viewing the computer screen. Thus, the keys that are mapped to these activities are configured as parallel keys. As such, each can be activated along with other parallel keys as well as other serial keys.

Any combination of keys can be configured to function as parallel keys. In fact, <u>all keys have the default status of "parallel" unless otherwise selected as serial or event keys.</u> Configuration of keys as parallel-, serial- or event-keys will be discussed later in this document.

Now press the "phone" key again to record the end of the conversation. Along with the auditory feedback you will notice the black tab disappear, as well as the ending of accumulation of time in the "phone" cell in row "Total Time On."

Take a moment to play around with different combinations of parallel keys.

Event keys:
These keys record discrete events.

A visitor walks in to the office. The ";" key at the extreme right of the screen is used to record that event. Press that key; notice that this time the black tab on the key lights up only momentarily. The ";" key — labeled the "visitor" key in this configuration — has been configured as an *event key*. This mapping reflects our interest in knowing how many visitors enter the office, not how long they stay.

<u>Event keys are used to catalog the occurrence of behaviors and events at a certain point in time; event keys record time and frequency of occurrence, not duration</u>. Thus, each press of an event key creates a time-stamped record of that event in the data file. As is the case for the serial and parallel keys, any and all of the keys may be configured to function as event keys, according to your own observational requirements.

Press this key several times and note that in the statistics box only the frequency count accumulates.

Sound:
The sound of a key-press can be turned off and on.

We have pointed out that one of the sources of feedback for a key press is an audible "click." While this may be useful for "blind" typing of inputs, it may be too distracting in situations in where the observed subjects are able to hear the click. To turn off the key-press sound, select Sound from the Settings menu. To turn the sound back on, re-select Sound from the Settings menu.

Notice that the check mark disappears when Sound is turned off and reappears when it is turned on.

Undo:
Use this command to erase the most recent key-press.
[Z]

Suppose you accidentally hit a key or you press a key in anticipation of an event that does not subsequently occur. Hold down the key and press "Z" to undo the last key-press; alternatively, you may select Undo from the Edit menu. Notice that the black tab on the last key activated (denoted by the highlighted bar in the statistics box) disappears, as does the data associated with that key-press in the statistics box. <u>The undo command is used to erase the most recent key-press</u>.

Note: If you explicitly save data with the "Save Data" command (see Save Data section), you cannot undo the last input performed before saving the data. Also, you cannot undo an "undo" command.

Replace:
Use this command to replace one key -press with another.
[*shift +* new key]

Suppose you observed the subject typing on the computer but accidentally pressed the "write" (D) key instead of the "type" (F) key. Hold down the shift key and press "F." Notice that the black tab on the "write" key disappears, while the black tab on the "type" key lights up.

Note also that in the statistics box, the frequency and duration counts have switched from the "D" key to the "F" key. <u>The Replace command allows you to instantly replace one key-press with another.</u> Whatever time was accumulated in the Total Time count of the first key (as a result of the last key-press) will be added to the Total Time count of the second key.

Pause key: *The recording session may be temporarily paused, then continued.* [*delete* key]

Let's pause here to review what we've done. Select the PAUSE button on the screen (the one that looks like a stop sign) or just press the "delete" key. Pausing A.C.T. stops the session clock and stops accumulation of time for all categories that are selected as "on" when the pause began.

The duration of the pause will be reflected in the data file in two forms: "pause on" and "pause off" are recorded on two separate lines, each of which has the real-time clock-stamp. Additionally, the "pause on" line displays the time on the session-clock when pause was selected, and the "pause off" line displays the time the session-clock would have shown had paused not been selected, i.e., [time-at-pause-on] + [duration-of-pause].

Restart (Start): [space bar]

Now select the START button or press the space bar. Note that the clock resumes its activity and all keys that were selected before the pause have resumed their accumulation of time in the statistics table.

Pause the program and restart it several times. Notice that a white tab on the *delete* key provides feedback of key selection, as does the very salient fact that the clock has stopped. Also note that whichever keys (or observational categories) were selected before the pause remain selected throughout the pause (though they do not accumulate time) and continue being on once START has been selected again.

Windows

A.C.T. enables you to open three different windows during a recording session. These windows allow you to enter text comments and to view statistics tables describing the data that has been logged up to that point in time. We will now open and view each of the three windows.

Note: The recording session continues in the background while a window is open. Pressing any key will close the window. If you observe a change in activity and thus have to press one of the active keys — e.g., a key that is part of the configuration, or a function key such as the "delete" (pause) key — the window will close and *the program will record and/or respond to whatever key was pressed.*

Statistics window:
Descriptive statistics.
[D]

This window allows you to view several statistics that describe the data collected up to the point at which the window was opened. The list includes, for each variable, the following measures: frequency, total time on, % time on, average time on, SD time on, minimum time on, median time on, maximum time on, and average time between on.

The statistics window may be opened through the Windows menu in the menu bar at the top of the screen or by holding down the key and pressing "D." Once the window is open, pressing any key will close the window.

Remember that pressing a key that serves a function in the running configuration will not only close the statistics window but will also activate that function.

Probabilities window:
Transition probabilities.
[P]

To open this window, hold down the key and press "P," or select Probabilities from the Windows menu. The probabilities window displays a matrix describing the observed probability of transitioning from one observational category to another. The numbers represent the probability of a first-order transition from a category on the left-hand column of the table to a category in the row at the top of the table (see "From" and "To" labels).

Remember that pressing a key that serves a function in the running configuration will not only close the data window but will also activate that function.

Comments window: *Comments and text entered at any time will be saved in the data file.* [⌘T]

Let's suppose that at this point during the session an unexpected event takes place. Perhaps the subject's computer breaks down, or they sit on their desk — some event or behavior that is important, yet does not have a particular key assigned to it. <u>The Comments window allows you to record text notes at any time during the session.</u>

Simply hold down the ⌘ key and press "T," or select "Comments" from the Windows menu. You may now enter up to three lines of verbal comments with automatic word wrap. Although the clock looks stopped on the display, the program continues to run. Your comments will be inserted into the data file, along with one time stamp identifying the precise time at which you opened the comments window and another time stamp identifying when you pressed the '*return*' key — or selected "Enter Comments" — to exit the comments window.

Remember that pressing a key that serves a function in the running configuration will not only close the data window but will also activate that function.

Save, Reset and Quit

Now that you know the essentials of recording data and opening windows, we turn to the three functions you need to know in order to complete the picture from the data collection perspective.

Save data:
Data is saved in the data file named at the session's start.
[S]

Data is automatically saved upon quitting the program or when resetting the program with a new data file name. You may also save the data file manually during the session or at its end: Simply select Save Data from the File menu, or hold down the key and press "S."

Each time you save data while A.C.T is running, a summary of descriptive statistics and transition probabilities will be printed in the data log section of the data file. Note that these summaries only reflect the data accumulated at the time data was saved, and thus will have different values than the final summaries at the end of the data file.

Go ahead and save the data, using either the menu or the key combination. If you save the data while in session, the program will continue running and collecting data normally.

Reset:
Resetting the session will reset the clock and all of the accumulated data records.
[*-delete*]

Hold down the key and press "delete." You will be asked if you are sure you want to reset the session. For now, in order to avoid resetting the program, press "Cancel."

Selecting "Reset" will return the clock to zero as well as erase all data that appears in the statistics window. If you start running the session again after "Reset" was selected, the new data will write over the old data in the data file, and no record will remain of the first session.

Note: As a safety feature, the program will ask you at each stage of the "Reset" routine whether you are sure about resetting and replacing the existing data file. You will also be prompted to name a new data file if you want to save data in the original file.

Quit:
This command terminates the session, closes the program, and saves the data.
[Q]

You are now ready to quit the program and to learn about configurations and data files. Select "Quit" from the File menu, or hold down the key and press "Q" to terminate the session, save your data to file, and exit the program. You've most probably noticed that the application has quit. Notice, however, that you now have a new file called "demo" in your folder. This is the data file created by you during this demonstration session.

Data Files

Viewing data files: In order to view the data you have just collected using Microsoft® Word© (or your default text processor), double-click the "demo" data file, or open it through the File menu on the menu bar. You may be asked to select a converter, with the "Text" option as the default. Select "OK" *(return).*

You will now see your data file in an unformatted text layout. You may immediately start reading the file. In order to more easily read the statistics tables at the bottom of the file, however, you will need to select the entire file (Select All from the Edit menu, or press -A) and make two adjustments to the document's format: 1) Choose a non-proportional font, such as Courier or Monaco, size 9 point; and 2) Set the left and right margins to 0.5" (using the arrows on the ruler bar, or through Page Setup in the File menu).

IMPORTANT: Data files should be saved in the original "Text Only" format in order to allow for post data processing routines (see "Processing Data Files," p. 17). If you want to save the data file as a Microsoft® Word© (or other formatted) document, make a copy of the text-only data file before doing so.

- **Document Setup and Summary Information:** At the top of the data file you will see a set of instructions that will serve you as a reminder for the above formatting instructions, followed by summary information displaying your file name, date and time of data collection, total recording time, total pause time, and total time of recording session.

```
!
!    ** IMPORTANT! IMPORTANT! IMPORTANT! IMPORTANT! **
!       To read or print this file in MS Word:
!       ------------------------------------
!         1) Select all of the text in the file
!         2) Set the font to a non-proportional font
!            like Monaco or Courier, 9 pt.
!         3) Set left and right margins to 0.5 in.
!            using Page Setup
!
!
!
!          Data File: 'demo'
!        Date & Time: Thursday, December 23, 1993  12:58:37 PM
!     Recording Time:   0:17:41.32
!        Paused Time:   0:01:13.08
!         Total Time:   0:18:54.40
!
!-----------------------------------------------------------------
```

- **Data Log:** Under this header you will find the Data Log. Here is where your time-stamped inputs are displayed, in chronological order. Data in the log is organized in four columns presenting, from left to right: 1. Key type and action; 2. Session time; 3. Real time; 4. Key label and action. Actions are coded as "+" for ON and "-" for OFF.

Here is a sample Data Log:

```
!     ==================================================
!     v                                                v
!     v                   Data Log                     v
!     v                                                v
!
#+    0:00:00.00     04:10:12
S+    0:00:24.25     04:10:36    "sit"+
A+    0:00:32.80     04:10:45    "stand"+
S+    0:00:44.50     04:10:56    "sit"+
A+    0:00:45.42     04:10:57    "stand"+
S+    0:00:49.65     04:11:02    "sit"+
L+    0:00:54.67     04:11:07    "phone"+
K+    0:01:00.80     04:11:13    "search.file"+
K-    0:01:18.78     04:11:31    "search.file"-
L-    0:01:38.02     04:11:50    "phone"-
F+    0:01:38.77     04:11:51    "type"+
J+    0:01:40.07     04:11:52    "read.screen"+
;-    0:01:56.47     04:12:08    "visitors"-
@+    0:02:03.45     04:12:15
@-    0:02:16.63     04:12:28
! Comment began at 0:02:53.85 and finished at 0:03:04.36
! This is where text comments are added
@+    0:03:02.77     04:13:28
@-    0:03:05.78     04:13:32
S-    0:03:02.78     04:13:32    "sit"-
F-    0:03:02.83     04:13:32    "type"-
J-    0:03:02.90     04:13:32    "read.screen"-
#-    0:03:02.95     04:13:32
!
```

Several symbols appear in certain rows in the first column:

- #+ This is A.C.T.'s symbol for "Start Session."
- #- This is A.C.T.'s symbol for "End Session."
- @+ This is A.C.T.'s symbol for "Start Pause."
- @- This is A.C.T.'s symbol for "End Pause."
- ! This precedes any entry that is not a data record, such as a text comment.

Since the Data Log is formatted to display one line per each key press, long observational sessions may generate data files that are several pages in length. For this reason, at the end of the Data Log you will once again see the data file summary information seen at the top of the program.

- **Configuration Setup:** Under the summary information you can see the Configuration Setup. This describes the observational category defined for each key and the particular function — e.g., event key, serial key — allocated to different keys. Remember, since all keys function as parallel keys by default, the program only lists those keys which were specifically defined otherwise.

```
!     =================================================
!     v                Configuration Setup              v
!     v                                                 v
!
! Configuration File: 'DEMO.A.C.T.'
!
!  Configuration uses Both hands.
!  'A' --> "stand"
!  'S' --> "sit"
!  'D' --> "radio"
!  'F' --> "type"
!  'G' --> "read.paper"
!  'J' --> "read.screen"
!  'K' --> "search.file"
!  'L' --> "phone"
!  ';' --> "visitors"
!  Serial keys = ( 'A', 'S' )
!  Serial keys = ( 'G', 'J' )
!  Event keys = ( ';' )
```

- **Statistics Summary:** Below the Configuration Setup, you will see statistic tables describing your total set of data. These tables are identical to the tables you saw earlier when you opened the "Data" and "Probabilities" windows.

The first table describes, for each variable, the following measures: frequency, total time on, % time on, average time on, SD time on, minimum time on, median time on, maximum time on, and average time between on.

The second table is a transition matrix listing the probability of first order transitions from one observational category to the other; enter the appropriate row and go across to the appropriate column to find the probability of transitioning "From" one behavior or event "To" another.

- **Text Comments:** At the bottom of the data file, you will see a summary of all text comments entered during the session. These are redundant with the comments listed individually in the Data Log and are grouped together for your convenience.

IMPORTANT: Data files should be saved in the original "Text Only" format in order to allow for post data processing routines (see "Processing Data Files," p. 17). If you want to save the data file as a Microsoft® Word© (or other formatted) document, <u>make a copy of the text-only data file before doing so.</u>

Customizing Your Own Configuration

A.C.T. was designed to allow for easy configuration of its interface, including labeling of keys, definition of key types (parallel, serial or event), and preferred mode of key layout (left handed, right handed, or both hands). By default, if you open A.C.T. by double-clicking on the A.C.T. icon, the keys are arranged for input with both hands, all key labels map to the key's letter, and all keys are in the parallel mode.

Open A.C.T. by double-clicking on the A.C.T. program icon. Enter any name you choose for the data file (Note: Do not use demo.1, demo.2, or demo.3 — these will be used later in the tutorial). You are now looking at the default configuration. The following instructions will take you through the different options that are available for customizing the configuration to your own needs.

Key layout:
The layout of displayed keys can be changed to allow typing with both or either hands.

Before you start labeling individual keys, you need to decide on the general layout of the on-screen keys. Would you like to type your inputs with both hands? Will you need to take notes with one hand while entering data with the other? Are you left handed or right handed? A.C.T. was designed to allow you to customize the key layout to your needs.

As you can see, the default layout is for entering data with both hands. You may change the layout of the keys by selecting "Type with Left Hand" or "Type with Right Hand" from the Settings menu, or by pressing -L or -R to select a left- or right-handed layout, respectively. Selecting (-B) will return the layout to the both-hands setting.

Select different layouts to familiarize yourself with this capability. Leave the configuration in whichever layout you prefer.

Note: The keyboard layout can only be changed before data collection has started.

Map keys mode:
In this mode, you define key function and label, and customize the clock.
[M]

To enter the mapping mode, you may either select "Map Keys" from the Settings menu, or hold down the key and press "M." As you will see, the Start, Pause and Reset keys have been disabled, indicating that you can <u>not</u> run a data collection session in this mode. Note that the clock window has changed its shading: this informs you that the clock settings may also be changed.

<u>In all, three things may be modified in the Map Keys mode:</u>
- Key labels
- Key functions
- Clock settings

Key labels:
Key labels may be changed to reflect different categories of observation.

There are two ways to input label names:
Press the *tab* key. Notice that the label for the "A" key has been highlighted. You may now write text into this label. When you've finished entering your label name, pressing *tab* again will highlight the next label (letter "S"); pressing *enter* will remove the cursor from the "A" label window. Repeated presses of the *tab* key will move the highlight through all label boxes. When the last (right-most) label is highlighted, a press of *tab* will highlight the first (left-most) label again. Pressing Shift-Tab will highlight the previous label.

Place the cursor on the label box for "A" and double-click: the box is highlighted. You may now write text into this label. When you've finished entering your label name, move over to the next label (letter "S") and double-click to highlight it. You may now write text into this label. This same procedure can be used to label all the keys.

Key functions:
Keys can be defined as parallel, serial, or event keys.

By default, all keys are defined as parallel keys.
To define serial keys:
Using the cursor and holding the shift key down for multiple selections, select those keys which you want to group as serial, i.e., mutually exclusive keys. You may also hold the cursor key down and drag the mouse to select multiple keys, as you would to select multiple objects/text in other Apple Macintosh applications.

When all related keys have been selected, open the Settings menu and select "Define Serial Controls" or hold down the key and press "G. " When you collect data using this configuration, this group of keys will act serially.

Note: You may define more than one group of serial keys. Once you have defined one group, simply click the cursor on another set of keys you wish to define as serial.

To define event keys:
Using the cursor, select the key(s) to be event keys. From the Settings menu, chose "Define Event Controls" or hold down the key and press "H. " When you collect data using this configuration, these keys will act as event keys.

"Undefine" key functions:
To return serial or event keys to parallel function, select Undefine Controls from the Settings menu.
[U]

When modifying an existing configuration or to correct a mistake while creating a configuration, you can Undefine individual and groups of keys that were previously defined as serial or event keys according to the following instructions.

Make sure you are in the "map keys" mode by either selecting "map keys" from the Settings menu or by holding down the key and pressing "M."

Using the cursor, select the key or group of keys to be undefined. Note that for serial key groups, selecting any one of the keys will highlight all the keys associated with that grouping.

Either select "undefine controls" from the Settings menu or hold down the key and press "U."

Setting the clock:
The session clock may be preset to match any other clock.

Sometimes you may want to synchronize the session clock with a particular clock, such as a time stamp on a video tape. Changing the clock setting from the default setting of 0:00:00.00 is done by individually changing each one of the segments (hours, minutes, seconds, or one-hundredths of seconds).
To reset the session clock:

Place the cursor on the particular segment you wish to change (hour/minute/second/100th). Double-click to highlight the segment and enter the desired setting.

Note: For obvious reasons, you may not enter numbers higher than 24 in the hour segment, numbers higher than 60 in the minute and second segments, and numbers higher than 100 in the 100ths segment.

Saving the new configuration:

Anytime you exit the "Map Keys" mode after changing or creating new labels or key functions, A.C.T. will prompt you to save the new configuration. We recommend that you always take this opportunity to name and save your configuration by selecting "yes" when prompted.

Opening pre-defined configurations:

Three ways to start A.C.T.:
1. Click on program icon, then select "Open Configuration" from the File menu.
2. Click on configuration icon.
3. Drag and drop configuration icon into program icon.

Processing Data Files

A.C.T. allows you to manipulate the date three additional ways, two of which allow you to combine separate files into a larger file (append and merge) and one which provides you with added statistical analyses (concurrency analysis).

Appending or merging data files allows you to do two things:
- You may take multiple files and append them sequentially into one long file; A.C.T. will subsequently provide you with the statistical analysis of the resultant file.

- You may have two or more observers collecting data simultaneously, each using his or her own computer and focusing on different activities — in effect, each operating a different subset of keys from the same configuration. Their data files may be merged to create one comprehensive file that includes all observations. The same can be done when transcribing a video recording, where one performs multiple passes over the same segment, each time creating a data file that describes different activities, in effect, using a different subset of keys from the same configuration.

Concurrency analysis allows you to :
- Examine the concurrence of different combinations of activities and events from a single file, or from an appended or merged file.

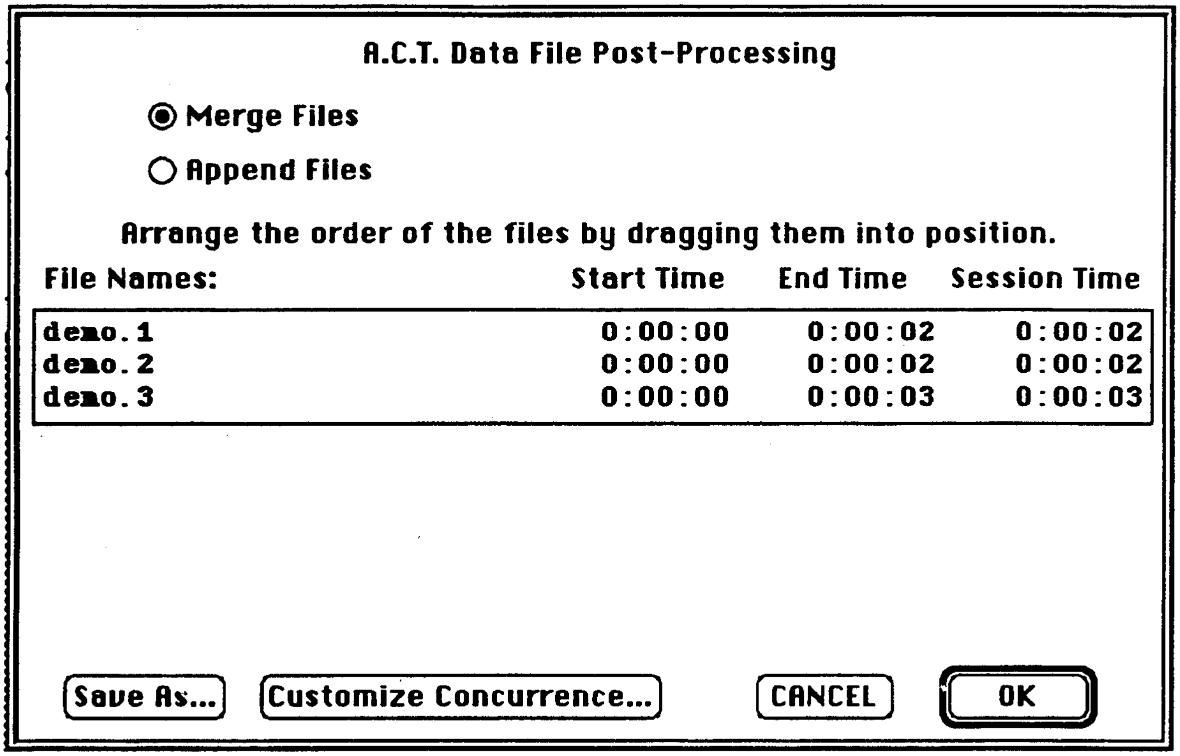
To Perform Data File Post-Processing:

To perform any one of the post-processing data file manipulations, you must follow the next steps:
1. Select those files that you would like to process.
2. Select the configuration that was used to create those files.
3. Drag and drop the selected files and the configuration together on to the A.C.T. program icon.

We have provided you with three data files — demo.1, demo.2, and demo.3 — with which you can learn about A.C.T.'s data-file processing functions. All these files were previously created using the DEMO.A.C.T. configuration. Select the three data files and the configuration file, then drag and drop them onto the A.C.T. program icon.

IMPORTANT: Data files should be saved in the "Text Only" format in order to allow for post data processing routines (see "Processing Data Files," p. 17). If you want to save the data file as a Microsoft® Word© (or other formatted) document, make a copy of the text-only data file before doing so.

At this point, the A.C.T. Data File Post-Processing window will open.



Appending files:
This process allows you to create one long file from several shorter files that were recorded in a sequence.

Select the Append Files button in the top left-hand corner of the window.

Note: The two fields that open in the right hand corner when you select Append Files are currently not functional. Subsequent versions of A.C.T. will allow you to use these buttons for greater control of data file processing.

When you append files, you are creating a chain of files which are connected "head to toe." A.C.T. allows you to select the particular order in which you want to append the files.

Notice that all the files you had chosen to manipulate appear in the A.C.T. Data File Post-Processing window. The order in which they first appear is determined by Session Time (from shortest to longest file). The files will be appended in the order in which they appear, with the top file first and the bottom file last.

To change the order in which the files appear in the window:
"Drag and drop" the line of text corresponding to each file to the desired location in the sequence.

Merging files: *This process allows you to create one file from two or more files that were recorded at the same time.*

Select the Merge Files button in the top left-hand corner of the window.

When merging files, A.C.T. consults the session clock to produce a file that combines all the activities and events according to the time at which they were recorded. For example, if one file has event A at 00:01 and event B at 00:08, and the other file has event C at 00:05, the merged file will have event A at 00:01, event C at 00:05, and event B at 00:08.

IMPORTANT:
• You can only merge files that were recorded using the same configuration.

• You can only merge files in which different keys from the same configuration were pressed, i.e., the same key (activity) cannot appear in more than one file.

• You can not merge files which "split" groups of serial keys. That is, members of serial key groups can only be activated within the same file.

Concurrency analysis: *This procedure produces two new data tables which present statistics regarding the occurrence of combinations of activities.*

One of the analysis tools that A.C.T. provides you with allows you to look at the concurrence of different activities and events. You might want to know: How often did the person talk on the phone while standing? How much time did the person write while sitting?

A.C.T. will print out all single activities and all possible combinations of two or more activities that were recorded concurrently in the session. (Notice that the "null" set of "no activity" is included in this listing.) Simply select a single file and its configuration and "drop" them into the program icon. When the A.C.T. Data File Post-Processing window opens, just click OK.

These data are presented following the statistical summaries described above (DATA FILES) in the form of two tables.

Description of concurrency tables:

!All single and concurrent activities, sorted by total duration
!(Listed single keys and combinations ONLY)

This table lists all recorded activities, including single activities, mutually exclusive combinations of keys, and the "null set." For example, the row that describes the concurrency of activities X and Y includes only those times when X and Y alone were activated; if activity Z came on while X and Y were on, the data is included in another row, namely, in the X and Y and Z row. The list is ordered by total duration, from longest to shortest.

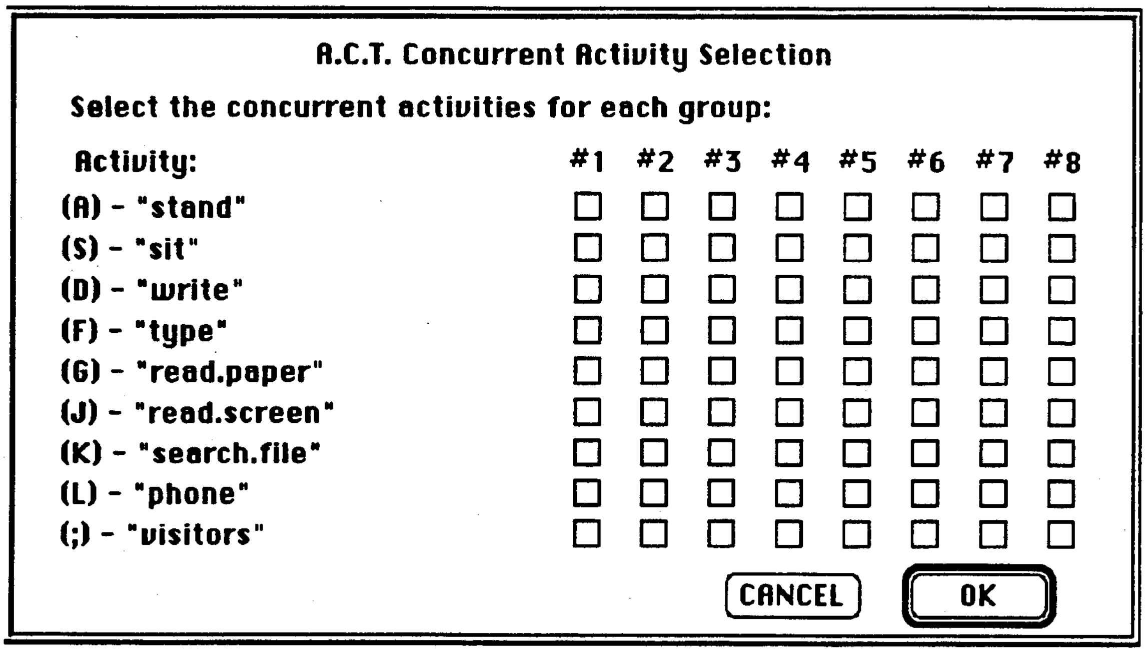
!All concurrent activities, sorted by total duration
!(Listed key combinations, REGARDLESS of other concurrent activities)

This table lists all possible combinations of 2 or more recorded activities. For example, the row that describes the concurrency of activities X and Y includes all times that X and Y were on at the same time, regardless of what other keys were activated at that same time. The list is ordered by total duration, from longest to shortest.

Customizing concurrency analysis:
You can define particular activities that interest you and create your own concurrency tables.

Remember that since <u>all</u> possible combinations of activities are presented in both default tables described above, the particular combinations which interest you will be included in this table. However, customizing the concurrency analysis will produce two tables which present only those combinations that interest you, thus freeing you from the need to search through what may be very long " !All single and concurrent..." and "!All concurrent activities..." tables.

To perform this customized concurrency analysis, you need to select the Customize Concurrence button in the A.C.T. Data File Post-Processing window. A new window will open: "A.C.T. Concurrent Activity Selection"



You may select up to 8 customized combinations of activities. Combinations of activities are described in columns. You define the specific combination(s) by selecting the desired combination of boxes which correspond to the list of activities on the left hand side. Once you have defined one or more combinations, click OK. Click OK again in the A.C.T. Data File Post-Processing window. In addition to all data tables (see DATA FILES section above), and in addition to the two concurrency tables described earlier in this section, the Concurrent Activity Table section of the data file will have two new tables:

"!All single and concurrent activities, selected by user
!(Listed single keys and combinations ONLY) "

This table presents only those unique activities, and/or combinations of activities, which you had selected in the customization of concurrency analysis.

" !All concurrent activities, selected
!(Listed key combinations, REGARDLESS of other concurrent activities) "

This table presents all the possible combinations of those activities which you had selected in the customization of concurrency analysis.

Summary

We have put much thought and time into making A.C.T. intuitive to operate. While we hope that the above instructions provide concise answers to questions you have, we believe that most questions can be answered simply by playing with the program. We encourage you to explore A.C.T.'s functions and capabilities in your daily surroundings: study your partner's activities as he/she cooks in the kitchen, try to find patterns in dialogs you hear on TV shows, or analyze the strategies employed by your favorite sports team. Remember that observational data collection depends primarily on the observer; A.C.T. is merely a tool, the utility of which will be defined by your choice of context and manner of application.

As these lines are written, we are aware of constraints and limitations inherent in our design. We intend to further refine this program and to add functions and capabilities that it does not currently provide. To this end, we depend on your feedback and inputs. Please send your comments to the address provided on the cover page, and keep in touch to receive our program updates.

We truly hope you enjoy A.C.T., and that in the course of its use, you find it a versatile and productive tool.

Acknowledgements

We would like to thank Dominic Wong for his wonderful programming talent and seemingly endless patience. Thanks also to Andrew Rice for help in debugging the program. This program was written with the support of NASA Ames Research Center, Aerospace Human Factors Division, FLR Branch; Sterling Software, Western Aerospace Laboratories, Inc. and Interface Analysis Associates.

Key Commands

Function	Key command	Description
Comments	T	The Comments window allows you to record text notes at any time during the session.
Create Data File	N	New data files can be created with this command.
Define Event Controls	H	Event keys are used to catalog the occurrence of behaviors and events at a certain point in time; event keys record time and frequency of occurrence, not duration.
Define Serial Controls	G	Serial keys are used to catalog behaviors and events that are mutually exclusive — only one can occur at any given time.
Map Keys	M	In this mode, you define the function of each key and give each key a label.
Open Configuration	O	Use this command to open previously-defined configurations.
Pause	*delete*	Pausing A.C.T. stops the session clock and stops accumulation of time for all categories that are selected as "on" when the pause began. Note: Pause can also be performed via a button on the display
Probabilities	P	The Probabilities window displays a matrix describing the probability of transitioning from one observational category to another.
Quit	Q	This command terminates the session, closes the program and saves the data.
Replace	*shift*-[new key]	The Replace command allows one to instantly replace one key press with another.
Reset	-*delete*	Resetting the session will reset the clock and all accumulated data records. Note: Reset can also be performed via a button on the display
Save Data	S	Data is saved in the data file named at the start of the session.

Function	Key command	Description
Start (restart)	spacebar	Hitting the space bar will (re)start your data collection session. Note: Start can also be performed via a button on the display.
Statistics	D	The Statistics window allows you to view several statistics that describe the data collected up to the point at which the window was opened.
Type with Both Hands	B	Selecting this option will configure the A.C.T. interface to two-hand typing.
Type with Left Hand	L	Selecting this option will configure the A.C.T. interface for one-hand typing using the left hand.
Type with Right Hand	R	Selecting this option will configure the A.C.T. interface for one-hand typing using the right hand.
Undefine Controls	U	Use this command to undefine previously-defined event and serial keys.
Undo	Z	The undo command is used to erase the most recent key press.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 1994	Contractor Report

4. TITLE AND SUBTITLE
Activity Catalog Tool (A.C.T.) v2.0 User Manual

5. FUNDING NUMBERS
NCC2-486

6. AUTHOR(S)
Leon D. Segal and Anthony D. Andre

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Western Aerospace Labs, Inc.
16111 Mays Avenue
Monte Sereno, CA 95030

8. PERFORMING ORGANIZATION REPORT NUMBER
A-94046

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NASA CR-177634

11. SUPPLEMENTARY NOTES
Point of Contact: Sandra Hart, Ames Research Center, MS 262-3, Moffett Field, CA 94035-1000
(415) 604-6072

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified — Unlimited
Subject Category 54

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report comprises the user manual for version 2.0 of the Activity Catalog Tool (A.C.T.) software program, developed by Leon D. Segal and Anthony D. Andre in cooperation with NASA Ames Aerospace Human Factors Research Division, FLR branch. A.C.T. is a software tool for recording and analyzing sequences of activity over time that runs on the Macintosh® platform. It was designed as an aid for professionals who are interested in observing and understanding human behavior in field settings, or from video or audio recordings of the same. Specifically, the program is aimed at two primary areas of interest: human-machine interactions and interactions between humans. The program provides a means by which an observer can record an observed sequence of events, logging such parameters as frequency and duration of particular events. The program goes further by providing the user with a quantified description of the observed sequence, through application of a basic set of statistical routines, and enables merging and appending of several files and more extensive analysis of the resultant data.

14. SUBJECT TERMS
Activity analysis, Observation tool, Behavior catalog software

15. NUMBER OF PAGES
28

16. PRICE CODE
A03

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified		

THE MAIN DIRECTIONS OF CVR DATA ANALYSIS DURING THE ACCIDENT/INCIDENT INVESTIGATION

A. S. Belan
Interstata Aviation Committee, Moscow

It is well known, that the CVR data is an essential source of information for the air safety investigator, as it is often the only recorded source of human performance information.

So, the CVR data analysis is the obligatory one, and is to be made by the field team on the accident site.

However, as the practical experience shows, the special laboratory research is required in some cases. First of all, it is occurred, when the recorder is badly damaged or the CVR data is needed to be defined more accurately. The main stages of this research are presented on the scheme (Appendix 1). As you can see, it is traditional enough.

The main directions of the laboratory research of the aural & sound data are announced in Appendix 2.

The verifying of the results of the listening through include:

— verification of the conversation content;

— verification of the sources;

— verification of timing.

To analyze the above items, different methods of assessment & estimation are utilized.

In order to obtain the additional information about the circumstances of the accident, the special laboratory research include:

— aural communication analysis of the cockpit conversation;

— speech analysis for the evaluation of the actual functional (psychophysical) condition of the crew;

— analysis of the sound situation in the cockpit for the assessment of the warning situation.

So, it is necessary now, to make the detail observation of the above directions (see also the Appendix 3).

The aural communication peculiarity analysis contain:

— indentification of the disturbances in aural communication reception and transmission, identification of the causes of such kind of disturbances and its result

— the research of the peculiarities of the intra-cockpit conversations.

Such methods as the psycho-linguistic method (contect-analysis) & acoustic analysis of different sources are used for this purposes.

Speech is certainly one of the most reach source of information about the condition of the speaker. This is also confirmed by the practice of the accident/incident investigation. The experience of the radioconverstion analysis shows, that a lot of problems, which are important one for the evaluation of the crew condition in flight can be solved with this kind of analysis.

Proceedings of the International Aerospace Congress *IAC* ' 94, (Edited by M. Liberzon), Belan, A.S., The main directions of CVR data analysis during the accident/incident investigation, *pp. 156-159, (1995)*,with kind permission from International Aerospace Congress.

So, this problems are:

— psycho-emotional stress dynamics & degree of its intensity (wording normal stress, increased stress, emotional stress);

— degradation of the influence of the negative effects in flight (for example hypoxy, acceleration, vibration & so on);

— condition of the static physic load (including overloads and great strength to the control units).

The applicated complex method include the utilization of the psycholinguistic & acoustic method.

The acoustic (noise) background suggest the indentification of :

— sound warning and alarm signals of the aircraft warning system;

— sound effects of the various bouard systems and units;

— operating engine noise changing;

— sound effects of the sturcture failures and so on.

To achieve the solution of this problems, the special acoustic methods were proposed to use.

The laboratory research of the CVR data requires good theory, update equipment & more over it is requires the excellent personal, which must have good command of language, as well as psychology, physiology & acoustic. Due to the particularity of acoustic research in the accident/incident investigation, the methods & facilities from the other spheres of industry are not useful fot he tasks of the accident/incident inveastigation. Therefore, it is significant to develop the theoretical ways of such kind of research, as the methods of practical operating, too.

According to the actual need of decreasing of the processing the CVR data, the experts of the Scientific Technical Center of the Commission for Flight Safety of the Interstate Aviation Committee created the complex program of the acoustic research on the base of IBM-compatible PC. This program allows to fulfill the following kind of analysis;

— auditing analysis;

— oscillographic (it contains the opportunity to choose & to zoom any content of the oscillographic record);

— analysis of the spectrum, in "frequence - intensity" coordinates (summary spectrum and cuts);

— spectrographic analysis, in "frequency - intensity - time" ("visual speech").

Although this program provides:

— the main useful signal filtration (including filtration for the low frequency, high frequency and other types of filtration);

— the reverse of the content of the oscillographic record (in order to produce the reverse listening of the speech content);

— speech timing as for the separate speech contents as for the full record.

The utilization of this program technically provides the converstaion analysis creation for all above mentioned problems and acoustic research direction.

In conclusion I should like to invite all specialists, who are interested for the cooperation in order to produce a new stage in the acoustic research and to exchange with the experience. Thank You!

Attâchment 1

The Main Stages Of CVR Data Analysis During The Accident/Incident Investigation

Evaluation of the CVR condition

Preparation of the CVR recording medium for reproduction

Reproduction & listening through the CVR magnetic tape

Transcription of the CVR tape

Identification of the sources

Timing of the communication in the cockpit

Synchronisation of the communicaton in the cockpit with flight data recording

Laboratory research of the CVR magnetic tape transcription

Studies, verifying the results of the listnening through

Special additional studies

Documentation of the analysis results

Attâchment 2

The Main Directions Of The Laboratory Pesearch Into The Aural And Sound Data, Recorded On The CVR Magnatic Tape

Verified results of listnening through

Verification of the communication content

Varification of the sources

Varification of timing

Special studies of the additional information

Aural communication analysis (information reception-transmission)

Speech analysis for the evaluation of functional condition of the crew-members

Analysis of the sound situation in the cockpit

The Main Contents & Methods Of Special (Additional) Studies

Analysis of the aural communication	Analysis of the speech to evaluate the dynamics of the functional condition	Analysis of the sound situation in the cockpit
The main tasks: — identification of disturbances in aural information reception and transmission — identification of the cuases of such disturbances — peculiarities of the intracockpit communication — peculiarities of air-to-ground communication	The main tasks: — psycho-emotional stress dynamics — evaluation of the degree of psycho-emotional stress — degradation of the functional condition — influence of hazardous and emergency situations (hypoxy, accelerations, vibrations, static loads)	The main tasks: — identification of the sound warning signals — identification of the sound effects of various systems & instruments operation — evaluation of the operating engines noise changes — identification of sound effects of structure failure and decompression
Methods used: — linguistic analysis — acoustic analysis	Methods used: — psycho-linguistic analysis — acoustic analysis	The method used: — acoustic analysis

APPENDIX B

ACOUSTIC PROCESSES

Editor's Note:
The manuscripts contained in this appendix are reprinted with written permission from the authors and publishers. We thank them for their help and support in making this report a more comprehensive repository of the methods and metrics used in voice communications.

On the contribution of instance-specific characteristics to speech perception

A. R. Bradlow, L. C. Nygaard, and D. B. Pisoni

Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405, U.S.A.

1. INTRODUCTION AND BACKGROUND

The role of variability in the listener's interpretation of the speech signal has been the topic of extensive research, and in general, it has been treated as a source of "noise" to be separated from the meaningful, abstract, symbolic units of speech [1,2]. For example, the general approach of many studies of speech acoustics has been to perform various measurements of speech sounds as produced by a small number of talkers in various phonetic and/or prosodic environments, e.g. [3-5]. The data are then used to derive generalizations about the nature of speech sounds and their contextual variation, which can then be used to investigate the acoustic cues to the related perceptual contrasts. An explicit assumption of this approach is that the variability inherent in the speech signal presents an "obstacle" to the listener that needs to be removed, or "stripped away", from the signal to facilitate perception of the underlying abstract linguistic units. Accordingly, the driving force behind this general research agenda has been the specification of the principles that underlie the observed variability in the speech signal so that it can be perceptually "undone" by the listener.

In contrast, our theoretical approach treats variability of the speech signal as a useful source of information that, though separate from the linguistic message, is available to listeners at all stages in their interpretation of the speech signal [6-8]. For example, this approach predicts that listeners will be sensitive to inter-talker differences; and that, rather than removing this source of variability from the signal as a consequence of perceptual analysis, listeners use this information as a basis for identifying talker characteristics that can aid in the interpretation of the linguistic message. Accordingly, in our acoustic analyses of sentences produced by multiple talkers we have deliberately avoided averaging across many talkers to derive summary generalizations about speech production; rather, we focus on inter-talker differences and try to correlate these differences with differences in listener responses. In general, our approach contrasts markedly with the traditional, "abstractionist approach" to speech because we focus on instance-specific variation, as opposed to the traditional emphasis on instance-independent generalizations about idealized, abstract symbolic forms [9,10].

In keeping with this general theoretical orientation the research presented in this chapter is motivated by three observations regarding variability in speech

In *Levels in Speech Communication: Relations and Interactions*, (Edited by C. Sorin, J. Mariani, H Meloni, & J. Scheontgen), Bradlow, A.R., Nygaard, L.C., & Pisoni, D.B., On the contribution of instance-specific characteristics to speech perception, pp. 13-24 (1995), with kind permission from Elsevier Science B.V., Amsterdam, The Netherlands.

perception. First, we observe variability in the intelligibility of different sentences across many talkers and listeners. Second, we observe variability in the intelligibility of different talkers across many sentences and listeners. And third, we observe variability in the perceptual strategies used by different listeners in learning to identify the voices of different talkers, and in their use of this talker-related information in speech perception. In other words, we observe that some sentences are more intelligible than others, that some talkers are more intelligible than others, and that some listeners make better use of instance-specific information in speech perception than others. The findings reported here represent an attempt to identify some of the specific utterance-, talker-, and listener-related correlates of speech perception.

Two sources of data provide the basis for our analyses. The first set of data come from a talker variability database of 100 Harvard sentences produced by 20 talkers (ten females and ten males) giving a total of 2000 recorded sentences [11]. This database also includes intelligibility scores for each sentence and talker. These scores were obtained from listening tests in which 200 listeners (ten per talker) transcribed each of the 100 sentences. Examination of these intelligibility data revealed considerable variability in the intelligibility of individual sentences and individual talkers.

The second set of data comes from a talker identification study [12], in which listeners were trained over a period of several days to identify the voices of ten talkers (five females and five males). The stimuli were recordings of isolated monosyllabic words produced by the ten talkers; nineteen listeners were trained over a nine-day period to identify the talkers by name. On the tenth day, subjects participated in two test phases: the first was a talker identification task in which subjects were required to explicitly identify the now "familiar" voices producing a new set of words; the second was a speech intelligibility task in which subjects identified a new set of words produced by either the old, familiar talkers or by a new set of ten unfamiliar talkers [12]. The results of this study provide information about the relationship between talker distinctiveness (that is, talker identifiability) and talker intelligibility, as well as data concerning the variability in the performance of different listeners in these two types of perceptual tasks.

Taken together, the results from analyses of the talker variability database and the talker identification study provided us with a rich set of data that we used to explore instance-specific correlates of speech intelligibility.

2. UTTERANCE-RELATED CORRELATES OF SPEECH INTELLIGIBILITY

We begin with an analysis of the specific sentence-related characteristics that correlate with variability in sentence intelligibility. The intelligibility tests using the 100 Harvard sentences from the talker variability database showed that the sentence intelligibility scores across all talkers and listeners ranged from 54% to 98% correct transcription, with a mean and standard deviation of 87.7% and 8.65%, respectively. In order to examine the sentence-related correlates of this variation in intelligibility, a set of high-intelligibility sentences was selected for comparison with a set of low-intelligibility sentences. The high-intelligibility set consisted of the 14 sentences with greater than 95% correct transcription; the low-intelligibility set consisted of the nine sentences with less than 75% correct transcription. All Harvard sentences have one clause consisting of five keywords and any number of additional function words. Accordingly, these sentence intelligibility scores are based on a scoring criterion which counts as correct only those transcriptions in which all five keywords are correct. Since all of these sentences are similar with respect to clause structure, our comparisons of the sets of high- versus low-intelligibility sentences focused on characteristics such as sentence length and the lexical characteristics of the individual keywords.

Our first finding in comparing the high-intelligibility sentences and low-intelligibility sentences was that the high-intelligibility sentences have fewer words on average (7.2 versus 8.2 words per sentence, $p(21)=0.03$). This count of words includes all words in the sentences, even though the sentence intelligibility scores are based on the correct transcription of only the five keywords in each sentences. The results suggest that the number of words surrounding the keywords has an effect on the overall sentence intelligibility: Words in longer sentences are more susceptible to error than words in shorter sentences. Furthermore, an examination of the repeated transcription errors for both set of sentences showed that almost all of the few errors on the high-intelligibility sentences can be traced to a low-level perceptual error, whereas for the low-intelligibility sentences many of the numerous errors can be thought of as higher-level "memory" errors. For example, a repeated error in the high-intelligibility sentences was found in the first word of the sentence, "Kick the ball straight and follow through", which was transcribed as "keep" more than once. Clearly, these two words are very close phonetically, as well as both being semantically compatible with the rest of the sentences. In contrast, a common error in the low-intelligibility sentences was the interchange of "strong" and "firm" in the transcription of the sentence, "The heart beat strongly and with firm strokes". In this case, the source of the error appears to be a memory confusion rather than a misperception. Thus, based on the error patterns exhibited by these examples it seems plausible that longer sentences have more transcription errors due to the higher memory load.

The second finding from our comparison of high- and low-intelligibility sentences examined the characteristics of the keywords. Across all Harvard sentences, the majority of the keywords were content words, that is, words that can be morphologically complex such as nouns, verbs, adjectives, and adverbs; however, in many cases the five keywords of a sentence included one or more function words, that is, words that are morphologically simplex such as auxiliaries, prepositions, pronouns, and demonstratives. A comparison of the keywords in the high- and the low-intelligibility sentences showed that the high-intelligibility sentences had a higher proportion of function keywords (21.4%) than the low-intelligibility sentences (11.1%). Since function words generally have a much higher frequency of occurrence in the language than content words, the higher proportion of function keywords in the high-intelligibility sentences leads to a higher mean word frequency for the keywords in the high- compared to the low-intelligibility sentences (1064 versus 152 occurrences per one million words of printed text, $p(113)=0.05$)[1]. Similarly, since function words are generally

[1] These word frequency counts are based on the Brown Corpus of printed text [13].

shorter than content words, the mean word length for the high-intelligibility sentences was shorter than for the low-intelligibility sentences (3.6 versus 4.1 phonemes per word, p(113)=0.025). These analyses suggest that overall sentence intelligibility is related to the mean word frequency *and* length of the individual words in the sentence, which are, in turn, related to their lexical status (that is, function versus content word).

Another difference between the high- and low-intelligibility sentences is related to the neighborhood characteristics of the keywords [14]. The "similarity neighborhood" of a word is the set of words that differ from the target word by a one phoneme substitution, deletion, or addition in any position [14]. The "lexical density" of a neighborhood is equal to the number of such neighbors, and the mean neighborhood frequency is the mean word frequency of all the words in a lexical neighborhood. Using these neighborhood characteristics we can describe a word as "easy" if it comes from a "sparse" neighborhood, and/or its frequency is higher than the mean neighborhood frequency of other phonetically similar words. Such a word has been shown to be more accurately and quickly identified than a "hard" word, that is, one that comes from a "dense" neighborhood, and/or does not occur with a higher frequency than its neighbors [14-16]. Using a computerized version of Webster's Pocket Dictionary, which is based on 20,000 entries, the neighborhood characteristics for the keywords in the high- and low-intelligibility sentence were found and analyzed[2].

As shown in Figure 1, for the high-intelligibility sentences the mean difference between keyword frequency (1140 per million) and mean neighborhood frequency (185 per million) is quite large (955 per million); whereas, for the low-intelligibility sentences the difference is 59 per million (209 - 150).

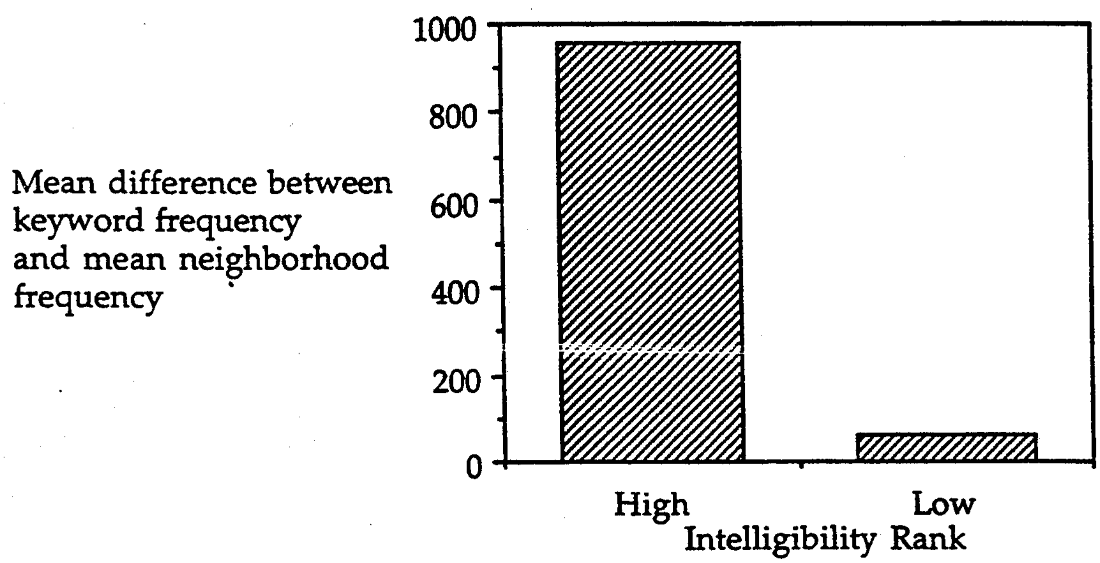


Figure 1. The mean difference between keyword frequency and mean neighborhood frequency for the high- and low-intelligibility sentences.

---

[2] Of the high intelligibility sentence keywords, 59 out of 70 (84.3%) appeared in this online dictionary; of the low intelligibility sentence keywords, 43 out of 45 (95.6%) were in this dictionary.

Additionally, a higher percentage of the keywords in the high-intelligibility sentences have higher frequencies than the mean frequency of the other words in their similarity neighborhood. In terms of mean neighborhood density, however, the high- and low-intelligibility sentence keywords come from equally dense neighborhoods (13.6 versus 13.3 neighbors per keyword). Thus, based on these analyses, the high-intelligibility sentences contain keywords that are more distinctive from their similarity neighborhoods in terms of word frequency, and they are therefore "easier" to recognize than the low-intelligibility sentence keywords. In other words, these words are more perceptually salient, and therefore less confusible with other phonetically similar words.

In summary, we have found that the number and nature of words that comprise a sentence have an effect on the overall intelligibility of the sentence, as measured by listener transcriptions. Specifically, words in longer sentences are more vulnerable to transcription errors than those in a shorter sentence. Additionally, the lexical and neighborhood characteristics of the individual words that comprise a sentence, such as word frequency and mean neighborhood frequency, correlate with its overall intelligibility. Specifically, on average, the high-intelligibility sentences have more function keywords than the low-intelligibility sentences, resulting in words that are generally more frequent and shorter in length. Furthermore, the keywords in the high-intelligibility sentences are perceptually more distinctive relative to other phonetically similar words in their lexical neighborhoods than the keywords in the low-intelligibility sentences. Earlier work has shown that such lexical and neighborhood characteristics are determining factors in the speed and accuracy of isolated word recognition [14-16]; the present results extend this finding to words in sentences by demonstrating that these same lexical characteristics play an important role in overall sentence intelligibility.

3. TALKER-RELATED CORRELATES OF SPEECH INTELLIGIBILITY

We now turn to a discussion of variability in talker intelligibility. The mean intelligibility scores across all 100 sentences for the 20 individual talkers in the talker variability database ranged from 81% to 93% correct transcription, with a mean and standard deviation of 87.9% and 3.1%, respectively. Many talker-related, or "indexical", factors might be expected to correlate with talker intelligibility, such as gender, overall speaking rate, dialect, fundamental frequency, vocal tract length and other details of speech production that can vary idiosyncratically from one speaker to another. In this section, our aim is to identify some of the talker-related factors that may affect speech intelligibility. We focus here on gender and overall speaking rate, as well as on two cases that examine talker-related details of speech timing in order to understand their perceptual consequences.

In a recent study of the TIMIT multi-talker database [17], Byrd [18] found that the prevalence of reduction phenomena, such as, increased speech rate, reduced frequency of stop releases, alveolar flapping, and vowel centralization were more prevalent among male speakers than female speakers. Based on this result, one might expect that the more carefully articulated speech of females would lead to higher intelligibility scores for females than for males. In fact, a comparison of the intelligibility scores for the female and male talkers in our database showed that the females have generally higher intelligibility scores than the males (89.4% versus 86.3% correct transcription, $p(18)=0.02$). Furthermore, all three of the talkers with intelligibility scores above 90% are female, and all three talkers with intelligibility scores below 85% are male. Thus, this correlation of gender and intelligibility in our database is consistent with the higher incidence of reduction phenomena for male talkers than for female talkers in the TIMIT database [15]. Taken together, these two results suggest that male and female speech differ in the precision of articulation, and that this difference has an effect on overall speech intelligibility. However, a direct connection between speech articulation and intelligibility for different talkers still remains to be made from the same source of data.

Overall speaking rate has been shown to be the primary factor that distinguishes carefully articulated speech from reduced speech, since many other reduction phenomena can be directly related to it [19-22]. Thus, we began by examining this factor in our attempt to explore the connection between reduction phenomena and overall speech intelligibility for male and female talkers. A comparison of the sentence durations for all 100 sentences for the three talkers with the highest intelligibility scores with those for the three talkers with the lowest intelligibility scores in the talker variability database revealed that, indeed, the former are longer than the latter (2054 versus 2008 milliseconds, $p(598)=0.03$). This suggests that overall speaking rate and intelligibility are factors that distinguish the most from the least intelligible talkers. However, we also found that the mean sentence durations for all ten males were longer than for all ten females (2155 versus 2085 milliseconds, $p(18)<0.001$), and that for all 20 talkers, mean sentence duration did not correlate with mean talker intelligibility ($r = 0.073$). Thus, although the most and least intelligible talkers in this sample can be distinguished by both gender and speech rate, when the whole set of speakers is included in the analysis, the correlations between gender and rate, and intelligibility and rate no longer hold. Furthermore, we found no evidence that sentence intelligibility and speaking rate correlate: there was no significant difference between the mean sentence durations for the fourteen high intelligibility sentences and the nine low intelligibility sentences (2125 versus 2149 milliseconds, $p=0.78$); and for all 100 sentences, mean sentence duration across all 20 talkers did not correlate with mean sentence intelligibility score ($r = 0.016$).

In summary, it appears that gender may indeed correlate with talker intelligibility; however, it is not immediately apparent that, for all speakers, this correlation is due to overall speaking rate. This result leads us to suspect that, although speech rate may play a role in determining the intelligibility of a talker (as shown by the rate difference between the three highest and the three lowest intelligibility scorers in our talker variability database), there are additional factors that can vary independently from overall rate and that contribute to overall talker intelligibility.

In order to investigate the fine-grained variability in timing details that may contribute to talker intelligibility, we present two cases of consistent listener errors that shed light on the perceptual consequences of some idiosyncratic timing differences between talkers. The first case comes from the phrase, "The walled town..." which was often transcribed by listeners as "The *wall* town ...". This error constituted 90% of the transcription errors for this sentence. In order to determine the acoustic factors that correlate with /d/ recognition in this phrase, various portions of the speech signal for each speaker were measured and then correlated with the rate of /d/ recognition for that speaker. The vowel-to-vowel durations, that is the portion between the /al/ of "wall" and the /a^u/ of "town," was measured from the point at which there was a marked decrease in amplitude and change in waveform shape as the preceding vowel-sonorant sequence ends, until the onset of periodicity for the following vowel. In almost all cases, this portion consisted of a single /d/-like closure portion and a single /t/-like release portion. Most talkers (18/20) did not release the /d/ and then form a second closure for the /t/. Furthermore, the /d/ closure portion generally consists of a period with very low amplitude, low frequency vibration, followed by a silent portion and then the /t/-like release burst and aspiration periods. Separate acoustic measurements of all of these components of the vowel-to-vowel period were taken, as well as the duration of the preceding /wal/ sequence.

Rank order correlations of these measures with the rate of /d/ recognition for each talker showed that the total vowel-to-vowel duration correlated quite highly with /d/ detection (Spearman rho = 0.702); however, an even higher correlation was found with the duration of voicing during the /d/ closure (Spearman rho = 0.744). In fact, this correlation between the absolute amount of voicing during the /d/ closure and the rate of /d/ detection for the individual talkers was stronger than any other proportional measure of this period. For instance, the rank order correlations between the proportion of voicing during closure to the total closure duration, and to the duration of the preceding word /wal/ were only -0.412 and 0.033, respectively. In other words, the duration of voicing during closure, in an absolute sense, appears to be the most reliable cue to the presence of a voiced consonant in this phonetic environment.

Inter-talker variability in voicing during voiced stop closure is a well-documented phenomenon in the production of American English, e.g. [23]; however, it is generally thought of as a less-reliable, secondary cue to stop voicing. The present correlation of the talker intelligibility data with the acoustic data provides a direct perceptual correlate of this source of variability and shows that listeners are, indeed, sensitive to this acoustic-phonetic variation, and use this information as a cue to the presence or absence of a segment.

The second case of a consistent listener error occurred in the phrase "the play seems", which was often transcribed by listeners as "the place seems". This error constituted 70% of all the transcription errors for that sentence. In this case, we examined the timing details of the acoustic signal in order to see what determined the syllable affiliation of the medial /s/. Measurements were taken of the duration of the /s/ (marked by the high frequency, high amplitude turbulent waveform) and of the preceding and following syllables (/plej/ and /simz/ respectively). Results showed that the absolute duration of the /s/ does not correlate very strongly with the rate of "play seems" transcription (Spearman rho = -0.254); whereas, when taken as a proportion of the /plej/ duration, that is, as a proportion of the preceding word, the rank order correlation with rate of "play seems" transcription is quite strong (Spearman rho = -0.653). In other words, the longer the /s/ relative to "play", the more likely it is to be syllabified by a listener as both the coda of the preceding word, and the onset of the following word. Thus, in this case the listeners appear to draw on more global information about the speaking rate of the talker in deciding on the placement of the word boundary (see [24,25]).

Furthermore, in this case, there appears to be a gender-related factor in the timing relationship between the medial /s/ and the preceding word, "play". Of the ten talkers with the shortest /s/ over /plej/ duration, and the highest percentage "play" transcription, eight are female; and, of the eight talkers whose renditions of this phrase were always correctly transcribed, seven were female. Thus, in this case, the female talkers as a group appear to be more precise with respect to controlling this timing relationship than their male counterparts. Although this case is not a matter of reduction (in fact, the correct form is shorter in duration), the apparent gender difference in speech production, which is, in turn, correlated with rate of correct transcription, is consistent with the hypothesis that the more carefully articulated speech of female talkers is also more intelligible. Moreover, this case provides an example that explains why overall speech rate is not the only, or even the primary, talker characteristic that determines talker intelligibility: finer acoustic-phonetic details of the timing relations between phonetic segments in an utterance make an important contribution to overall speech intelligibility.

4. LISTENER-RELATED CORRELATES OF SPEECH INTELLIGIBILITY

Information about the variability in speech intelligibility due to listener-related factors was obtained from the talker identification training studies, in which the listeners were divided into two groups based on their performance during training [12]. In this study, nineteen listeners were trained to explicitly identify by name the voices of ten talkers producing isolated, mono-syllabic words. By the ninth day of training, nine listeners were able to identify the talkers with greater than 70% accuracy; whereas, the remaining ten listeners failed to reach this level of accuracy.

Figure 2 shows the scaling solutions of the confusion matrices for the two groups of listeners on Day 1 (Figures 2a and 2b) and Day 9 (Figures 2c and 2d) of the training period[3]. On Day 1 of training both groups of listeners were effective at separating the female and male speakers along dimension one (DIM 1); and, for both groups at this stage, speaker M2 is distinctive in this dimension. However, within the male and female groups of speakers, the individual speakers are not very well distinguished along either of the other two dimensions for both the "good" and "poor" listener groups. By Day 9 of training, however, the "good" listener group appears to distinguish the female talkers along dimension three (DIM 3) and the male talkers along dimension two (DIM 2). In contrast, by the end of training the listeners in the "poor" listener group seem to have tried to use all three dimensions to distinguish each of the ten listeners, and, as a result

[3] These scaling solutions were generated from confusion matrices that counted the number of times listeners confused each voice with each of the other nine voices (see [26,27]).

they are less successful at the talker identification task than those in the "good" listener group. Thus, these scaling solutions demonstrate that listeners differ in the strategies they use to learn to explicitly identify different talkers, and that talkers differ in their distinctiveness. This finding raises two issues. First, does learning to explicitly identify the voice of a talker help in a word recognition task with words spoken by the familiar voice? And second, is talker distinctiveness related to overall talker intelligibility?

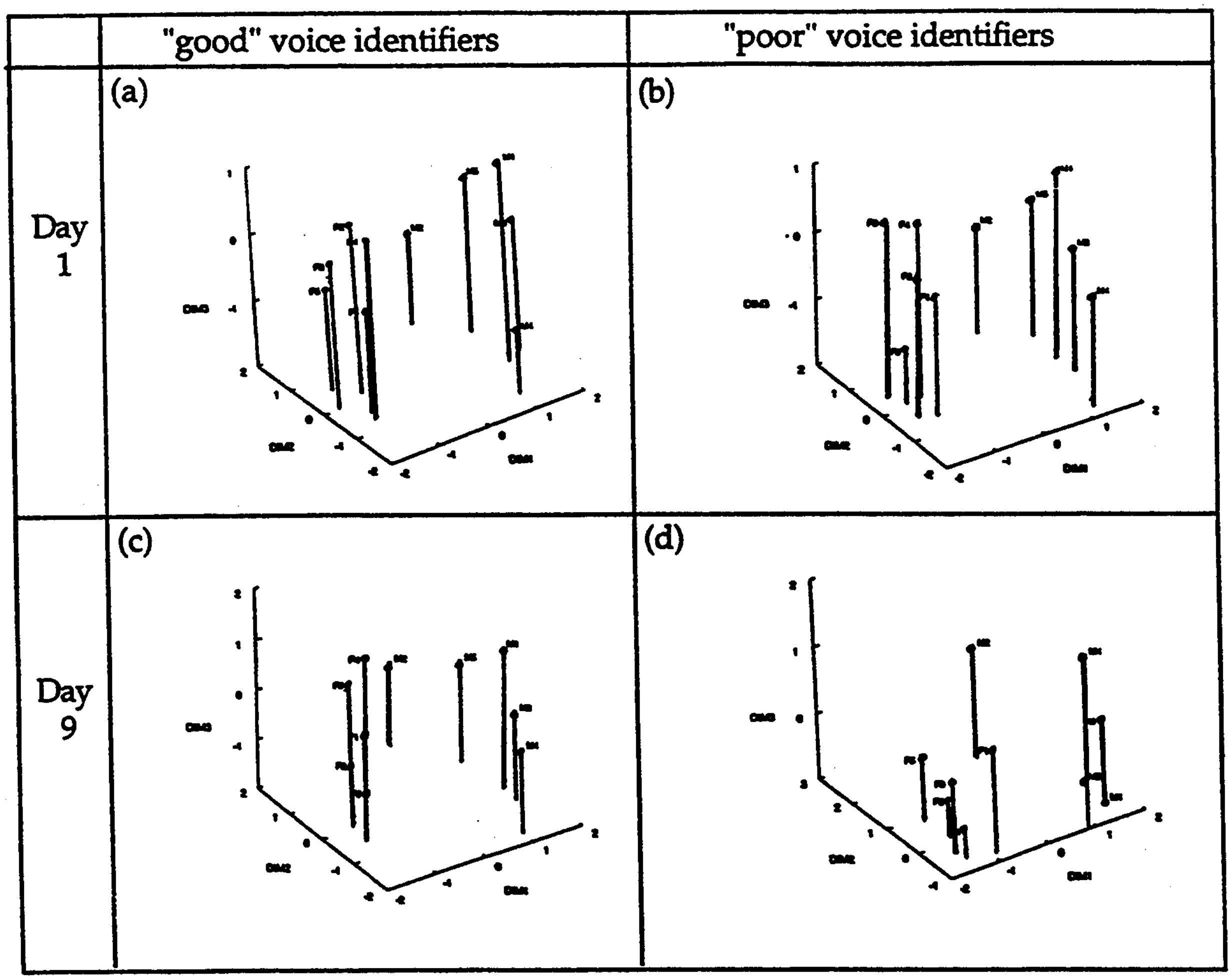


Figure 2. Scaling solutions for the listener data from the talker identification training studies: (a) Day 1 for the "good" listener group, (b) Day 1 for the "poor" listener group, (c) Day 9 for the "good" listener group, (d) Day 9 for the "poor" listener group (from [26]).

In response to the first issue, we found that in the test phase of the study, the "good" listeners showed an advantage in the word recognition task for novel words produced by familiar voices over novel words produced by unfamiliar voices; whereas, the "poor" listener group did not show any difference due to voice familiarity in the word recognition task. Thus, the "good" listeners apparently use their knowledge about a talker's voice such that their performance on a word recognition task is enhanced relative to the "poor" listeners. This result suggests that listeners differ in their ability to learn to identify talkers' voices and that these differences in perceptual learning do indeed affect speech perception.

We have seen that from the listener's point of view, individual voice identification and word recognition interact, producing an advantage in the recognition of novel words spoken by familiar voices relative to unfamiliar voices (see also, [27-29]). A related question is whether talker distinctiveness and talker intelligibility are correlated; in other words, is the most distinctive voice also the most intelligible voice? It is clear from the data in the talker variability database that some talkers are more intelligible than others. Furthermore, it is clear from the talker identification training study that some talkers' voices are more distinctive than others; for example, as shown in the scaling plots in Figure 2, Talker M2 is easily distinguished from the other nine talkers at the start of the training period by both the "good" and "poor" groups of listeners. However, the data from the talker identification training study indicate that although Talker M2 is the most easily identified across all listeners, this talker has the second lowest word intelligibility scores across all listeners and words. Furthermore, the overall rank order correlation for the ten talkers' identifiability and intelligibility scores is quite low (Spearman rho = -0.143), indicating that voice intelligibility and identifiability are not well correlated. Thus, it would appear that from both the listener's and the talker's points of view, individual voice identifiability and speech intelligibility are separate factors that, although not correlated, can interact to the extent that instance-specific characteristics are employed in the general processes of speech communication.

5. CONCLUDING REMARKS

The findings reviewed in this report suggest that the "indexical" [30] or "personal" properties of speech may play an important role in speech perception by placing constraints on phonetic and lexical interpretation. Human listeners apparently do not discard the fine instance-specific phonetic details that are encoded in the speech signal. As we have seen from two separate sets of analyses, these acoustic-phonetic details are preserved in memory and provide a rich source of information to assist in speech perception.

Specifically, the results of these investigations provide a clear demonstration of the relationship between variation in speech intelligibility and variation of the speech signal due to sentence- and talker-related characteristics. The results also show that the lexical and neighborhood characteristics of the words that comprise a sentence correlate with its overall intelligibility, implying that lexical characteristics that determine isolated word intelligibility operate at the sentence-level as well. Additionally, we found a correlation between inter-talker differences and overall talker intelligibility, suggesting that listeners are sensitive to the fine-grained acoustic-phonetic details that distinguish the speech of one talker from another, and that these differences contribute to a specific talker's overall intelligibility. Taken together, the correlation between word-level characteristics and overall sentence intelligibility, and the correlation between fine-grained acoustic-phonetic differences and overall talker intelligibility, demonstrate the important role that variability plays in controlling speech intelligibility. Thus, the pattern of results that emerges is one in which seemingly small, detailed effects are retained throughout the process of speech perception. The results indicate that, rather than being normalized to fit an abstract, idealized symbolic representation of the meaningful units of speech, these sources of low-level variability in the acoustic signal "propagate up" to higher levels of processing to modulate speech intelligibility.

The results of the talker-identification training study provide a direct demonstration of listener-related differences and the effect these strategies have on speech perception. The data also show that a listener's ability to learn to identify talkers' voices transfers to the recognition of new words produced by the familiar talkers. Thus, listeners apparently retain "talker-specific" information in memory and make use of this stored information in speech perception and spoken word recognition. This study suggests that speech perception is a "talker-contingent process," and that the talker-specific, indexical properties of speech may not be clearly dissociated from the abstract, linguistic properties; rather, listeners appear to be sensitive to *both* types of information in the speech signal, and that knowledge about a talker's specific vocal tract properties may assist in the perception of that talker's speech. We interpret these results as providing a demonstration of the contribution of instance-specific information to speech perception. Rather than viewing the inherent variability of the acoustic speech signal as "noise" that is somehow filtered out, or "normalized", by the processes of speech perception, we consider instance-specific variability as information in the stimulus that is directly encoded in the neural representation of speech, and is operative throughout the processes of speech perception and spoken word recognition.

REFERENCES

1. D. Shankweiler, W. Strange, and R. Verbrugge (1977). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, Acting and Knowing*. Potomac, MD: L. Erlbaum.
2. M. Studdert-Kennedy (1974). The perception of speech. In T. A. Sebeok (Ed.), *Current Trends in Linguistics.* The Hague: Mouton.
3. T. Crystal and A. House (1982). Segmental durations in connected-speech signals: Preliminary results. *Journal of the Acoustical Society of America,* 72, 705-716.
4. T. Crystal and A. House (1988). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America,* 83, 1553-1573.
5. V. Zue and M. Laferriere (1979). Acoustic study of medial /t, d/ in American English. *Journal of the Acoustical Society of America,* 66, 1039-1050.
6. D. Pisoni (1992). Some comments on invariance, variability and perceptual normalization in speech perception. In J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodge, & G. E. Wiebe (Eds.), *Proceedings 1992 International Conference on Spoken Language Processing.* Alberta, Canada: University of Alberta.

7. M. Sommers, L. Nygaard, and D. Pisoni (1992). Stimulus variability and the perception of spoken words: Effects of variations in speaking rate and overall amplitude. In J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodge, & G. E. Wiebe (Eds.), *Proceedings 1992 International Conference on Spoken Language Processing.* Alberta, Canada: University of Alberta.
8. J. Elman and J. McClelland (1986). Exploiting lawful variability in the speech wave. In J. Perkell and D. Klatt (Eds.), *Invariance and Variability in Speech Processes.* Hillsdale, NJ: L. Erlbaum.
9. L. Brooks (1978). Nonanalytic concept formation and memory for instances. In E. Rosch and B. Lloyd (Eds.), *Cognition and Categorization.* Hillsdale, NJ: Erlbaum.
10. L. Jacoby and L. Brooks (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. Bower (Ed.), *The Psychology of Learning and Motivation.* New York, NY: Academic Press.
11. J. Karl and D. Pisoni (1994) The role of talker-specific information in memory for spoken sentences. Poster to be presented at the 127th meeting of the Acoustical Society of America, Cambridge, MA.
12. L. Nygaard, M. Sommers and D. Pisoni (1994). Speech perception as a talker-contingent process. *Psychological Science* , 5, 42-46
13. F. Kucera and W. Francis (1967). *Computational Analysis of Present Day American English.* Providence, RI: Brown University Press.
14. P. Luce (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception, Technical Report No. 6,* Indiana University.
15. D. Pisoni, H. Nusbaum, P. Luce, and L. Slowiaczek (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication* , 4, 75-95.
16. P. Luce, D. Pisoni and S. Goldinger (1990). Similarity neighborhoods of spoken words. In G. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistics and Computational Perspectives.* Cambridge, MA: MIT Press.
17. V. Zue, S. Seneff and J. Glass (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication,* 9, 351-356.
18. D. Byrd (1992) Sex, dialects, and reduction. In J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodge, & G. E. Wiebe (Eds.), *Proceedings 1992 International Conference on Spoken Language Processing.* Alberta, Canada: University of Alberta.
19. M. Picheny, N. Durlach, and L. Braida (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research,* 28, 96-103.
20. M. Picheny, N. Durlach, and L. Braida (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research* , 29, 434-446.
21. M. Picheny, N. Durlach, and L. Braida (1989). Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *Journal of Speech and Hearing Research,* 32, 600-603.

22. R. Uchanski, K. Miller, C. Reed, and L. Braida (1992). Effects of token variability on resolution for vowel sounds. In M. E. H. Schouten (Ed.), *The Auditory Processing of Speech: From Sounds to Words.* New York, NY: Mouton de Gruyter.
23. L. Lisker and A. Abramson (1964) A cross-language study of voicing in initial stops: acoustical measurements. *Word* , 20, 394-422.
24. J. L. Miller (1987). Rate-dependent processing in speech perception. In A. Ellis (Ed.), *Progress in the Psychology of Language.* Hillsdale, NJ: Erlbaum.
25. J. L. Miller, F. Grosjean, and C. Lomanto (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica,* 41, 215-225.
26. L. Nygaard and M. Kalish (1994) "Modeling the effect of learning voices on the perception of speech. Poster to be presented at the 127th meeting of the Acoustical Society of America, Cambridge, MA.
27. D. Van Lancker (1991). Personal relevance and the human right hemisphere. *Brain and Cognition* , 17, 64-92.
28. D. Van Lancker, J. Kreiman, and K. Emmorey (1985). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. *Journal of Phonetics* , 13, 19-28.
29. D. Van Lancker, J. Kreiman, and T. Wickens (1985). Familiar voice recognition: Patterns and parameters. Part II: Recognition of rate-altered voices. *Journal of Phonetics* , 13, 39-52.
30. J. Laver and P. Trudgill (1979). Phonetic and linguistic markers in speech. In K. Scherer and H. Giles (Eds.), *Social Markers in Speech.* Cambridge, UK: Cambridge University Press.

Selected Topics in Forensic Voice Identification

Bruce E. Koenig
Federal Bureau of Investigation
Quantico, VA

Ongoing law enforcement operations throughout the world are continually capturing the voices of suspects with miniature transmitter/receiver systems, analog and digital on-the-body recorders, telephone intercept devices, and concealed room microphones. Since these recordings are normally utilized for investigative leads and/or legal proceedings, specific speakers must be accurately identified. Voice identifications that occur through self-recognition of one's voice, eye-witness information, surveillance logs, and the use of a person's name in the conversation are usually readily accepted. However, voice identifications that involve listening only and/or laboratory tests are often more difficult to evaluate accurately. To provide a better understanding of these voice comparison topics, two types of aural-only comparisons will be discussed, and an update on the spectrographic technique is included.

Aural Identification of Familiar Voices

Recognition of familiar voices is a daily occurrence for most people, as they identify spouses, children, coworkers, friends, and business associates after only a few words spoken over the telephone or by hearing them from an adjacent room. This process involves long-term memory, where recognition occurs through a prior knowledge of speech characteristics, including such attributes as accent, speech rate, pronunciation, pitching, vocabulary, and vocal variance (intraspeaker variability).

Some of the relevant scientific research and opinions that address the accuracy of identifying familiar voices include the following:

1. Researchers used 7 listeners who were familiar with the 16 chosen speakers through daily contact. The speakers had no pronounced speech defects or accents. Groups of two to eight speech samples of varying lengths were played back to the listeners, which resulted in an identification accuracy of better than 95% for samples lasting from about 1 to 2 seconds. Voice samples were also frequency restricted, but the results reflected only a limited loss of accuracy under conditions normally encountered in law enforcement investigations. In tests involving whispered speech, the duration had to be somewhat greater than three times longer than normal speech samples to obtain equivalent levels of identification (Pollack *et al.* 1954).
2. Sixteen listeners with no hearing losses, who had known the recorded 10 male coworkers for at least 2 years, were chosen. None of the 10 recorded individuals had either pronounced regional accents or speech abnormalities. When the listeners heard sentences of less than 3 seconds duration from the 10 coworkers, their median accuracy rate of identification was 98% (range of 92% to 100%). When only a disyllable (*e.g.*, mama) was spoken, the median accuracy rate dropped to 88% (range of 73% to 98%) (Bricker and Pruzansky 1966).
3. In a study of coworkers, recordings were made on different telephone lines of four women and seven men, each talking for 30 seconds to 1 minute on a neutral topic such as the weather. An additional recording was prepared of another male, who was relatively unfamiliar to most of the listeners. The recordings were arranged in a random order and played to 10 of the other coworkers, who were asked to identify the speakers. "All the listeners except one correctly identified all the 11 [coworkers]... The one listener who made an error... confused two speakers who were not well-known to him. Three of the 10 listeners knew [the eighth male, who was not a coworker], and correctly identified him. Of the remaining seven listeners, only two said that they could not recognize this speaker. Five listeners wrongly identified this speaker as..." another one of their coworkers. "It is worth noting that four of the five listeners who made the wrong identification were highly skilled, experienced phoneticians..." with doctoral degrees in the field (Ladefoged 1978). This experiment reflects a 100% identification rate for the coworkers' voices that were well-known to them and an overall average accuracy rate of 96% when the relatively unfamiliar voice was added.
4. Twenty-four individuals were asked to listen to speech samples of 24 coworkers (15 males and 9 females) whom they had known for several years and 4 speakers unknown to the listeners. The speech samples averaged about 30 seconds in length and contained at least 12 utterances of 2 to 4 words each. Listeners rated each coworker on a scale of very familiar to totally unfamiliar prior to the testing. They listened to the samples for as long as they wished and then rated their decisions as follows: (1) guessing, (2) fairly sure, or (3) very sure. Deleting the results of any voice rated totally unfamiliar to the listener, the results showed a 90.4% correct identification rate and 4.3% incorrect identification rate, with 5.3% who said they did not know the speaker. If the 5.3% are deleted, the correct identification rate is 95.4%. "This rate is probably fairly representative of situations where a limited vocabulary is required and can be expected to be even higher in informal conversations where more of the individual speaker's speech habits are present as cues for identification" (Schmidt-Nielson and Stern 1985).

Crime Laboratory Digest, Vol. 20, No. 4, Koenig, B.E., Selected topics in forensic voice identification, pp. 73-81, (1993). Reprinted with the permission of FBI Crime Laboratory Digest.

5. In an introduction to another research paper, the author states that the "Identification of speakers by their voices is a common experience. Most listeners have little difficulty in identifying the voices of familiar speakers over the telephone or on the radio. Recognition of voices of familiar speakers in the darkness or when the speaker is out of sight of the listener is also a common occurrence" (Compton 1963).

This research reflects that the identification accuracy rate for familiar voice samples lasting 1 second or longer ranged from 92% to 100% and averaged 95% to 100%. Samples recorded through the telephone or other limited bandwidth systems had little effect on accuracy. The effects of noise and loss of high frequency information were studied in another experiment (Clarke *et al.* 1966) which found that aural speaker identification was only slightly degraded when progressing from high-quality voice samples to typical investigative recordings. It is obvious from everyday experience and the cited research that identifying familiar voices can be an accurate method for identifying voices recorded in forensic applications, even with the limiting factors of noise and attenuated high frequencies.

Aural Identification of Unfamiliar Voices

Aural comparisons of unfamiliar voice samples rely on short-term memory. For example, a woman receives a number of different telephone inquiries regarding a classified advertisement. She then receives an obscene telephone call, and she tries to remember if any of the voices match. In a judicial proceeding, a judge and/or a jury may have to decide if a particular crucial comment on an investigative recording was spoken by the defendant, who readily admits to saying the other statements attributed to him on the transcript, or to someone else involved in the conversation. Examiners using the spectrographic technique, described later, play back the separate voice samples concurrently on separate devices or computer files with an electronic patching arrangement to allow rapid aural switching between them or by recording short phrases or sentences from each sample on the same recording (Voice Comparison Standards 1991).

The *de facto* study of unfamiliar voice comparisons (Clarke *et al.* 1966) determined the following:

1. Sentence length over the range of 5 to 11 syllables is not an important variable in identification accuracy.
2. Correct identifications decreased from approximately 90% to 80% when the signal-to-noise ratio (SNR) was reduced from 30 decibels (dB) to 0 dB.
3. Correct identifications decreased from approximately 88% to 78% when the frequency response was reduced from 4,500 hertz (Hz) to 1000 Hz.

Since most investigative recordings have a SNR of 10 dB to 40 dB and a frequency response of 2,500 Hz to 5,000 Hz, the range of expected correct identifications of unfamiliar voices would be 78% to 90%, with most identifications in the 78% to 83% range.

The use of expert testimony for aural identifications of unfamiliar voices provides no assistance to the court and/or to the jury. The notes of the advisory committee on Rule 901 of the Federal Rules of Evidence appropriately reflect this fact as follows: "Since aural voice identification is not a subject of expert testimony, the requisite familiarity may be acquired either before or after the particular speaking which is the subject of the identification..." (Federal Criminal Code and Rules 1991). Additionally, the voice comparison standards of the International Association for Identification (IAI) specifically state that it "... does not support or approve the use of... aural only expert decisions..." for voice comparisons (1991).

Spectrographic Comparisons

The spectrographic laboratory technique is the most well known and possibly the most accurate of the laboratory testing procedures presently available for comparing verbatim voice samples under forensic conditions. However, some scientists believe that aural identifications of very familiar voices are more accurate (Hecker 1971). The spectrographic technique has been described in numerous forensic and scientific publications, including an overview article published in the *Crime Laboratory Digest* (Koenig 1986). Therefore, a detailed explanation will not be rendered here; the following paragraphs provide a brief summary of the examination, a review of the new comprehensive standards passed by the IAI, and its status in government and private laboratories.

When properly conducted, spectrographic voice identification is a relatively accurate but not conclusive examination for comparing a recorded unknown voice sample with a suspect repeating the identical contextual information over the same type of transmission system (*e.g.*, a local telephone line). The examiner uses both the short-term memory process previously detailed and a spectral pattern comparison between identically spoken sounds on spectrograms. Figures 1A and 1B are sound spectrograms of different male speakers saying "salt and pepper." The horizontal axis represents time, divided into 0.1-second intervals by the short vertical bars near the top, and the vertical axis is frequency, ranging linearly from 80 Hz to 4000 Hz, with horizontal lines every 1000 Hz. The speech energy is reflected in the gray scale from black (highest level) to white (lowest level). The frequency range of the voice is analogous to the range of a musical instrument, where the lowest notes are at the lowest frequency and the highest notes at the highest frequency. The mostly horizontal bands of darkness reflect the vocal resonances and are called formants. The closely spaced vertical striations represent fundamental frequency (voice pitch) or the actual vibrations of the vocal cords. The spectrographic technique requires comparison of identical phrases between the voice samples, with a decision made at one of a number of confidence levels. The scientific support of this examination is limited, and the actual error rate under most investigative conditions is unknown. The research to date indicates that the technique has a certain error rate that is independent of examiner-induced errors, with errors of false elimination (the voice samples were

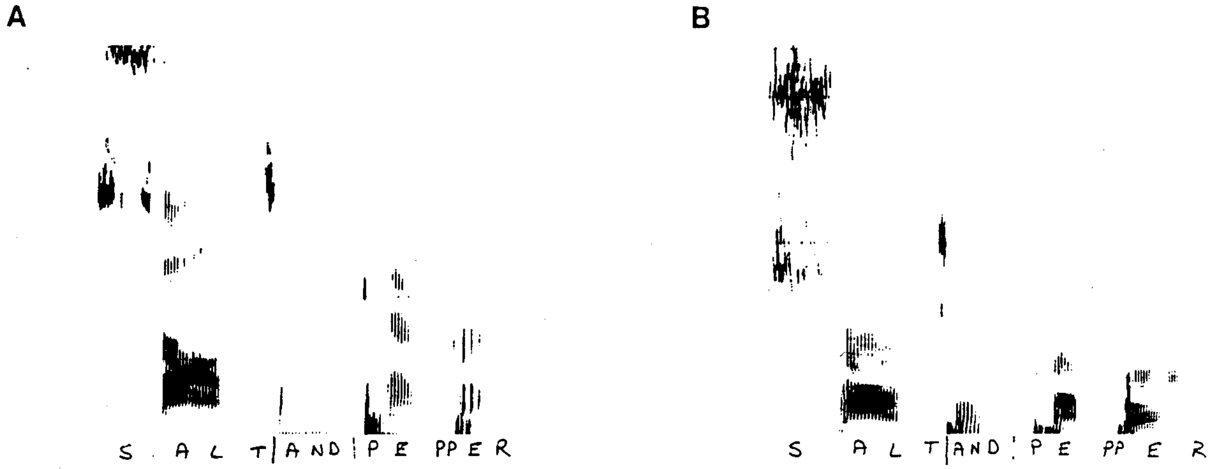


Figure 1 (A) and (B). Sound spectrograms of different male speakers saying "salt and pepper."

actually from the same person, but the examination found that they did not match) appreciably higher than false identification (the voice samples were actually from different persons, but the examination found that the samples matched).

In July 1991, the Voice Identification and Acoustic Analysis Subcommittee of the IAI passed and published its first set of comprehensive spectrographic voice identification standards. These requirements, which became effective January 1, 1992, for all certified IAI members, include examiner qualifications, evidence handling, preparation of exemplars, preparation of copies, preliminary examination, preparation of spectrograms, spectrographic/aural analysis, work notes, testimony, certification, and miscellaneous subjects. Table 1 lists the minimum qualifications for spectrographic examiners of the IAI and the FBI and updates a similar table published in an earlier issue of the *Crime Laboratory Digest* (Koenig 1986). Table 2 is another updated and expanded table from the same article concerning minimum criteria for spectrographic comparisons. Tables 1 and 2 and the previously published tables reflect that the upgraded IAI standards are now appreciably closer to the FBI's criteria. The FBI's standards require higher educational levels, more words for lower confidence decisions, enhancement procedures when needed, and a higher frequency voice range. The most important legal difference is the FBI's policy not to provide testimony on spectrographic comparisons due to the inconclusive nature of the examination and the unknown error rate under specific investigative conditions.

The use of the spectrographic technique since the mid-1980s continues to show a steady decline by both government laboratories and private examiners. As of mid-1993, the New York City Police Department and the FBI were the only government laboratories in this country regularly conducting these examinations. The private sector efforts were limited to less than a dozen part-time examiners. Professional meetings in the field have been sparsely attended, and no major spectrographic research is known to be under way. Problems still persist in the spectrographic voice identification field. Examples of these problems include the following: (1) separate sets of certified examiners making high-confidence decisions for both identification and elimination in the same case;[1] (2) individuals with no experience, training, or education in the voice identification discipline making conclusive decisions under oath in court; and (3) examiners testifying that an unknown voice is not the defendant's, although admitting their decisions are really inconclusive based upon accepted standards.

Table 1. Minimum Qualifications for Spectrographic Examiners of the IAI and FBI

Qualification	IAI	FBI
Education	High School Diploma	BS Degree
Periodic Hearing Test	Yes	Yes
Length of Apprenticeship	Usually 2 Years	2 Years
Number of Comparisons Conducted	100	100
Attendance at a Spectrographic School	Yes	Yes
Formal Certification	Yes	Yes

Table 2. Minimum Criteria for Spectrographic Comparison for the IAI and FBI

Criteria	IAI	FBI
Words Needed for Highest Confidence Level	20	20
Words Needed for Lowest Confidence Level	10	20
Affirming Independent Second Decision	Yes	Yes
Original Recording Required	Yes	Yes
Allows Testimony	Yes	No
Completely Verbatim Known Samples	Usually	Usually
Speech Frequency Rate	Above 2 KHz	Above 2.5 KHz
Accuracy Statement in Report	Yes	Yes
Enhancement Procedures, When Needed	Optional	Yes
Speed Correction of All Recordings	Yes	Yes
Track Determination of All Recordings	Yes	Yes
Azimuth Alignment Correction	Yes	Yes

Summary and Conclusion

Under investigative conditions, individuals can reliably identify voices that are well known to them, but the accuracy rate drops to approximately 78% to 83% when unfamiliar voices are compared to known voice samples. The use of expert witnesses does not improve the accuracy rate of aural only voice comparisons. The use of the spectrographic technique continues to decline, even with the establishment of new standards in 1992.

References

Bricker, P. D. and Pruzansky, S. Effects of stimulus content and duration on talker identification, *Journal of the Acoustical Society of America* (1966) 40:6:1441-1449.

Clarke, F. R., Becker, R. W., and Nixon, J. C. *Characteristics that Determine Speaker Recognition.* Technical Report ESD-TR-66-636, Electronic Systems Division, US Air Force, 1966.

Compton, A. J. Effects of filtering and vocal duration upon the identification of speakers, aurally, *Journal of the Acoustical Society of America* (1963) 35:11:1748-1752.

Federal Criminal Code and Rules. West, St. Paul, MN, 1991, p. 289.

Hecker, M. H. L. *Speaker Recognition: An Interpretive Survey of the Literature.* American Speech and Hearing Association, Washington, DC, 1971.

Koenig, B. E. Spectrographic voice identification, *Crime Laboratory Digest* (1986) 13:4:105-118.

Ladefoged, P. Expectation affects identification by listening, *Language and Speech* (1978) 21:4:373-374.

Pollack, I., Pickett, J. M., and Sumby, W. H. On the identification of speakers by voice, *Journal of the Acoustical Society of America* (1954) 26:3:403-406.

Schmidt-Nielson, A. and Stern, K. R. Identification of known voices as a function of familiarity and narrow-band coding, *Journal of the Acoustical Society of America* (1985) 77:2:658-663.

Voice comparison standards, *Journal of Forensic Identification* (1991) 41:5:373-392.

Note

1. Los Angeles Board of Civil Service Commisioners. Threat case decided March 25, 1992, in which three IAI examiners made an identification at a high-confidence level, while two IAI examiners eliminated the suspect.

Enhancement of Forensic Audio Recordings*

BRUCE E. KOENIG

Federal Bureau of Investigation, Engineering Research Facility, Lorton, VA 22079, USA

A detailed description of the Federal Bureau of Investigation's techniques to improve the intelligibility of investigative recordings is given, including equipment used, methodology, training of examiners, testimonial procedures, and evidence handling.

0 INTRODUCTION

Since the early 1960s, the Federal Bureau of Investigation (FBI) has been conducting examinations to improve the voice intelligibility of tape recordings produced by federal, state, and local law enforcement agencies, other federal organizations, and foreign governments. These submitted tapes, numbering in the tens of thousands, are often produced with miniature recorders or low-power RF transmitter/receiver systems in high-noise environments, and involve investigations of kidnapping, political corruption, drug trafficking, espionage, child pornography, presidential assassination attempts, and so on. This paper deals with the details of the FBI's methodology in conducting the examinations of these forensic recordings, where forensic refers to the application of science to the legal field, which includes criminal and civil investigations, presentation of evidence at court, and general assistance to the professionals in all aspects of the judicial system. The laboratory techniques have been developed based on scientific literature, experience, and the evolution of specialized audio devices. The following sections will cover forensic and studio recording differences, the equipment used, examination methods, examples of examinations, training of examiners, evidence handling, and testimony. Since this field is so empirically oriented, it will not be possible to describe every facet of the analysis process; however, the basic steps will be summarized with appropriate examples.

1 TAPES: FORENSIC VERSUS STUDIO

Most members of the Audio Engineering Society are involved or concerned with the production of very high-quality recordings, whether classical music, hard rock, or voice. Members regularly debate in print or informally such subjects as digital dithering, amplifier peak overload problems, loudspeaker design, and the best recording medium. In contrast, the forensic tape examiners of the FBI are concerned about whether a recording can be made understandable with the best possible playback system and state-of-the-art filtering. A "good" investigative recording might only have a 20-dB signal-to-noise ratio, a flat frequency response to 3 kHz, and some audible distortion, but still be completely understandable. Table 1 lists some of the more obvious differences between forensic and studio recordings. Tapes produced during law enforcement operations often are subjected to enhancement procedures to improve intelligibility for investigative purposes or introduction in courts of law, where the conversations can be understood by judges and juries with only a single playing during the judicial proceedings.

2 EQUIPMENT

The types of laboratory equipment and materials used to enhance forensic recordings can generally be categorized into the following classifications:

1) Standard professional analog and digital tape recorders
2) Logging tape recorders
3) Consumer-type tape recorders
4) Specially modified tape recorders
5) Fast Fourier transform (FFT) analyzers
6) Analog and digital filters
7) Analog and digital gain-reduction devices
8) Professional headphones
9) Digital audio storage devices
10) Professional amplifiers, cables, connectors, etc.
11) Microscopic and macrophotographic systems
12) Ferrofluids
13) Movable equipment racks

Fig. 1 shows a typical FBI laboratory setup.

* Manuscript received 1988 June 6; revised 1988 August

Tape recorders provide the means of accurately playing back law enforcement recordings and producing enhanced and direct copies for the contributor. Professional recorders are used for the playback of standard open-reel and cassette evidence tapes, and to produce laboratory and field investigative copies. Many of the units have been modified by the manufacturer for transport speeds as low as $\frac{15}{32}$ in/s. Consumer-type recorders are used for reproduction when professional decks are not available, such as miniature cassettes or 8-track cartridges. Logging recorders used by law enforcement agencies to record incoming telephone calls and police radio traffic normally operate at either $\frac{15}{32}$ or $\frac{15}{16}$ in/s and can have up to 60 channels of information recorded simultaneously for 25 hours. Since different manufacturers of these units use a wide range of mostly nonstandard track configurations, playback systems with time code readers have been purchased from various companies in the ¼-, ½-, and 1-in-wide tape formats. In addition, when these manufacturers upgrade with newer equipment, the time code and track configurations are frequently incompatible with the older equipment. The FBI has had some of its loggers modified for separate amplification of each channel, instead of the normally equipped summation amplifiers. Which brand and model of the professional, consumer, and logging recorders to purchase is usually decided by in-house testing.

Specialized recorders fall into three categories: old and obsolete devices, unique recorder formats, and

Table 1. Differences between studio and forensic recordings.

Characteristic	Studio	Forensic
Signal-to-noise ratio	60 dB+	Negative to 30 dB
Frequency response	20 to 20 kHz	100 Hz to 3–5 kHz
Distortion	Inaudible	1–10%
Wow and flutter	Inaudible	Inaudible to 1% rms
Equipment operator	Trained technician	Investigator
Microphone	Large professional	Miniature
Tape recorder	Professional analog and digital	Inexpensive to professional analog
Tape type	Best	Standard
Noise reduction	Yes or digital	Usually not used
Reverberation	Usually damped	High
Microphone-to-speaker distance	Close	Varies
Microphone location	Open	Hidden
Transmission system	Usually none	Often telephone or low-power RF

Fig. 1. Typical FBI enhancement setup.

standard playback units with custom-fabricated head stacks and modified electronics. Old and obsolete devices include wire recorders, magnetic dictation disk and belt units, and short-lived or specialized cassette and cartridge formats. As an example of their use, in 1981 the FBI was able to play back the Dallas, Texas, Police Department radio transmissions of November 22, 1963, which had been recorded on a Dictaphone Dictabelt and a Gray Audograph disk during the assassination of President John Fitzgerald Kennedy [1], [2]. Unique formats include four-in-line cassette recorders, very low-speed standard and miniature cassette formats, specialized on-the-body recorders, and others. Many of these playback units have been modified for improved fidelity and to add proper output connections. The last type of specialized device includes professional recorders with slow to medium tape speeds and custom-fabricated head stacks with modified electronics that allow adjustment of the reproduce magnetic heads to correct for alignment problems present on forensic tapes. The modifications are made with the cooperation of the recorder manufacturers and an independent tape head fabricator. As an example, the FBI has a top-of-the-line professional ¼-in open-reel tape recorder with tape speeds of $^{15}/_{16}$ through 7½ in/s with the original head stack removed and replaced with two separate playback heads (Fig. 2). The heads can be moved across the entire width of the tape and allow for azimuth adjustments of better than $\pm 6°$ from perpendicular. This reproduce unit is used to play back misaligned recordings that occur in such operational situations as "black box" cockpit voice recordings in major airplane crashes and when tape recorders are often moved from one investigative location to another.

The FFT analyzer is the focal point of the enhancement examination, since it provides the examiner with a continually updating visual representation of the recorded audio information. Without this device, non-automatic filtering is reduced to a purely aural evaluation that would rarely provide the optimal enhancement of voice intelligibility on forensic recordings. The analyzer graphically displays the parameters of frequency in the horizontal dimension and amplitude in the vertical, with Fig. 3 showing, as an example, a 800-Hz square wave from 0 to 5 kHz and an amplitude range of 1.0 V. A detailed description of FFT theory is not being provided, since many excellent sources are available on the subject [3]–[5]. The FBI uses a variety of analyzers, but most provide at least 800 lines of resolution, two separate channels, 4 kHz or better real-time bandwidth, linear and exponential averaging, frequency ranges up to 100 kHz, interactive cursor controls, plotter outputs, and high-resolution screen displays. In addition the analyzers provide a basic waveform display that can be helpful to the examiner with certain recorded noise problems.

Analog and digital filters utilized include professional bandpass, parametric, notch, comb, tracking, deconvolution, special, and, to a very limited extent, graphic equalizers. All of these devices and software supports are purchased or modified to operate principally below 7 kHz due to the band-limited nature of the investigative recordings received. Thus, for example, the separate parameteric filter modules for the high audio frequencies have been replaced with lower frequency ranges. The bandpass filter's low- and high-frequency settings are selectable in one-third-octave or narrower steps with a 24- to 48-dB-per-octave rolloff. The analog units are often based on a Butterworth filter design [6]. The parametric equalizers have user-adjustable center frequencies, bandwidths, and attenuation/gain controls and a separate in/out switch for each frequency band. The notch filters have adjustments for the center frequency and bandwidth, with Q values (center frequency $\div$ bandwidth) ranging from 1 to 1000. Octave, one-half-octave, and one-third-octave graphic equalizers are rarely used for intelligibility improvement, due to their limited resolution and the availability of other

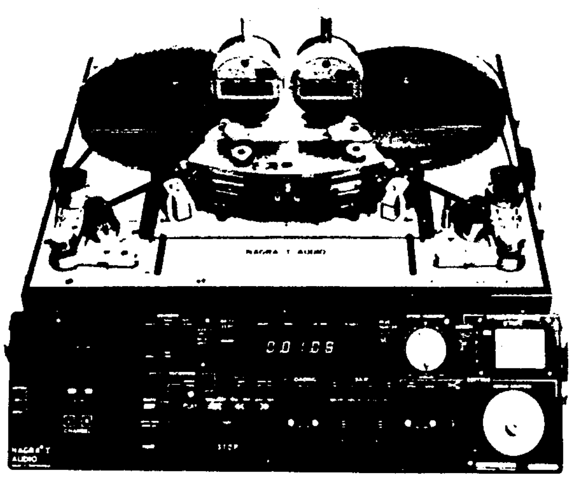


Fig. 2. Professional open-reel recorder with two specialized head stacks.

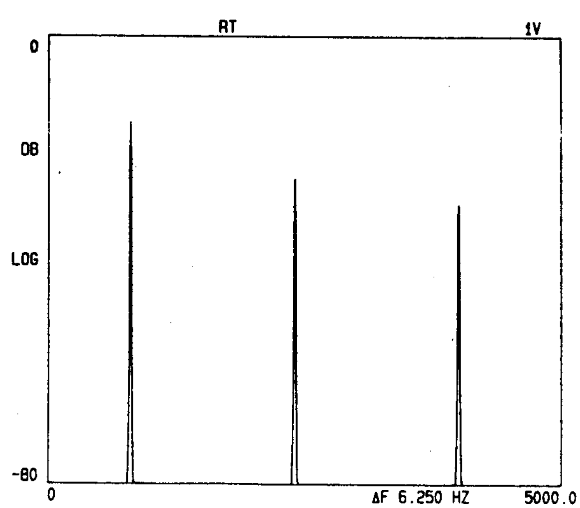


Fig. 3. FFT display of 800-Hz square wave (0–5-kHz range).

filter types, but they have some applications in spectrographic (voice print) comparisons [7]. The digital comb filters are 12–16-bit resolution devices that allow the attenuation of a discrete frequency tone and its harmonics. The fundamental frequency, number of harmonics to be reduced, and bandwidth of the notches are all user controlled.

The deconvolutional filter, which is rarely used in recording studios, is a digital processor with 12–16 bits of resolution that reduces the level of certain noises via an adaptive predictive deconvolution procedure. With a linear prediction algorithm, a transversal filter uses past values of the input signal to predict future audio information, and thus its effectiveness is highly dependent on the time correlation of components in the recording. For example, a pure sine wave is correlated, repeating itself every cycle, whereas random noise (that is, white or pink) is uncorrelated and voice information becomes uncorrelated in periods longer than a few hundred milliseconds. In Fig. 4 the transversal filter acts as a predictor with instructions from the adaptation processor and estimates the noise $\hat{N}$ slightly in the future. This estimate is then subtracted from the input signal, which contains the voice V and noise N components, producing output E, defined as $E = V + N - \hat{N}$. As the noise becomes more correlated ($\hat{N}$ approaches the value of N), the reduction in the nonvoice information by the device becomes larger [8]. The best digital deconvolution devices offer real-time filter orders of more than 5000 to handle reverberant recordings in large rooms, separate settings for high- and low-amplitude signals on the same recording, and adjustable filter size and convergence times. Additional digital filtering software of a specialized nature is run on a high-speed computer system with an array processor to handle one-of-a-kind problems and help develop algorithms that can be used in real-time, stand-alone processing devices.

The analog gain reduction/limiter devices are the types usually found in the better recording studios. The digital units have a "look ahead" feature that is especially important with sudden high-amplitude signals, such as recorded gunshots. All of these devices have variable compression ratios, attack and release times, and thresholds. Professional headphones, digital memory devices, amplifiers, movable equipment racks, and so on, need no further explanation.

Microscopic and macrophotographic systems or other types of magnification devices are combined with a ferrofluid to identify the track configuration and determine major azimuth misalignment. Low-power optical, photographic, and/or video display units in the range of 3 to 50 power and ferrofluids containing iron particles in the 0.2–1.5-μm size are used for many enhancement examinations. Fig. 5 shows a magnetic configuration sometimes encountered in forensic situations, where the track is offset to near the middle of the recording tape and the azimuth is badly misaligned. If played back on a standard cassette deck, the audio information would not have been heard, since it is recorded in the guard band between sides.

3 EXAMINATION

Enhancement examinations of original forensic recordings are experience-oriented analyses that defy easy quantification. Starting with the original recording, unless it has been destroyed, lost, or altered, the FBI examiner usually follows 12 basic steps that produce the most usable product for the contributor. The following discussion will explain these steps with detailed examples, but it is understood that a proper apprenticeship, as set forth later, is really the only way to properly grasp all the techniques. The 12 steps are:

1) Evidence marking
2) Physical inspection
3) Recorded track position and configuration
4) Azimuth alignment determination
5) Playback speed analysis
6) Proper playback setup
7) Overall aural review
8) Overall FFT review
9) Setup of enhancement devices
10) Copying process
11) Work notes
12) Reporting.

Before the examination begins, the submitted original evidence is marked for identification by the examiner using a permanent-ink pen with the assigned specimen number, laboratory number, and his or her initials. The FBI uses consecutive "Q" designations for individual recordings, that is, Q1, Q2, Q3, . . . , and a laboratory number that includes receipt date in the Laboratory Division's Evidence Control Center (ECC) and designators for the particular forensic section and examiner.

After the recording is properly marked for identification, a general physical inspection is conducted, as

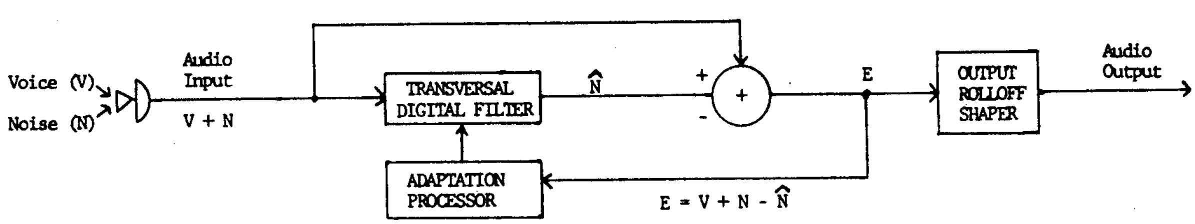


Fig. 4. Deconvolution filter design where $\hat{N}$ is the estimated noise. (Courtesy of Digital Audio Corporation.)

appropriate, of the housing, the reel, and the tape itself to ensure that all safety tabs have been removed, the housing or reel is not defective, the tape transports smoothly, and there are no obvious playback obstructions. If there are housing or reel problems, high-quality replacements are substituted and properly marked, and the defective housings or reels are returned to the contributor with the rest of the recorded evidence. Submitted recordings marked as copies are normally set aside until contact is made with the contributor to determine whether the submitted evidence is actually the original or not.

The recorded track position and configuration is determined by applying a diluted solution of a ferrofluid to the oxide side of a high-level recorded area on the tape. This is usually done by laying the tape on an absorbent surface, applying the ferrofluid with a small plastic squeeze bottle, and then slightly elevating one end of the surface until the solution evaporates. The developed tape is then placed below the lens of a low-power magnification system and a visual determination is made of the exact track configuration, for example, ¼ track, and its location on the tape. For example, a ½-track cassette recording that is severely offset from the tape edge would probably be best played back on the right channel of a ¼-track deck or a modified ½-track unit. The magnetic tape is then carefully cleaned with a freon-based cleaner to remove the ferrofluid residue.

General azimuth alignment examinations are often done visually with a close examination of the parallel magnetic striations in the recorded information at the time the record track is determined. Exact azimuth alignment is accomplished by adjusting the reproduce head on the playback unit for maximum high-frequency output using the FFT analyzer. Also, if the approximate frequency response of the recorded information is known, such as long-distance telephone conversation, the analyzer is used to measure the recorded frequency response of the evidential recording for direct indications of the loss of the higher frequencies. Table 2 reflects the signal loss at 20 min (one third of a degree) of azimuth misalignment at 3 kHz with various formats and transport speeds. It is readily apparent that the lower speeds combined with the relatively wider tracks on the forensic formats cause considerably greater losses for the same angular error [9].

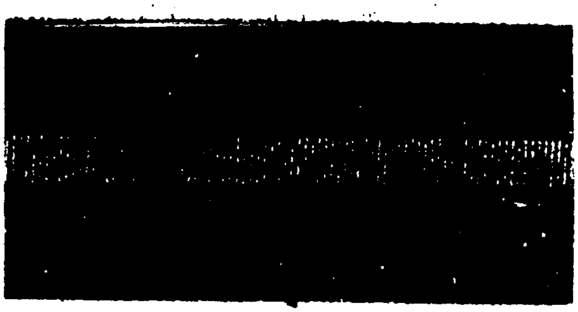


Fig. 5. Developed magnetic track with severe offset and azimuth misalignment.

The proper playback speed of the recording is determined by measuring any known discrete tones on the tape, again using the FFT analyzer. The most commonly encountered tones are 60 Hz and harmonics from ac power line leakage, various telephone signaling sounds (Touch-Tone, busy, and so on) and specialized power and RF components (for example, the 400-Hz tone present on many aircraft cockpit recordings). Fig. 6 shows an analyzer display of the 0–100-Hz range showing that the 60-Hz tone is actually at 63.5 Hz, as determined by playing the recording on a calibrated laboratory recorder, reflecting that the recorder used to make the recording was running approximately 5.5% slow. The lower quality recorders often utilized in many forensic applications produce greater speed discrepancies than normally encountered in recording studio operations, and the speed error may also vary considerably over the length of the recording [10].

Using the information gleaned from the physical inspection, recorded track position and configuration, azimuth alignment, and tape-speed analyses, the playback unit is configured to allow the best audio output. For major track and azimuth alignment problems, recorders with specialized reproduce head stacks are used for proper positioning, and speed changes are corrected by variable-speed controls. The output vu-meter levels are noted if overly high or low.

An overall aural examination with headphones is next conducted of the evidence tape to determine the approximate length of the recording, tape speed, type of information (such as telephone conversation), and to generally categorize the problems limiting intelligibility. Normally both sides of cassettes and the entire length of standard open-reel tapes are reviewed, unless otherwise designated by the contributor. Intelligibility problems noted when listening are often separated into the following areas to show general characteristics, but this should not be considered a complete list of problems encountered.

1) *Nonlinear distortion.* This type of distortion results in the clipping or flattening of the higher amplitude information in the audio waveform and the production of odd and even harmonics in the frequency domain,

Table 2. Amplitude losses at 3 kHz for 20-minute azimuth misalignments using different tape formats.

Format	Track	Speed (in/s)	Loss (dB)
Standard cassette	¼	1⅞	0.7
Standard cassette	½	1⅞	5.0
Standard cassette	½	15/16	19.8
Microcassette	½	15/16	19.8
Open reel	full	7½	4.7
Open reel	½	1⅞	8.6
Open reel	½	15/16	13.3
Open reel	¼	3¾	0.6
Logging reel (1-in tape)	40	15/32	7.8

and can be caused by overdriven electronic components, saturated record levels, the use of poor-quality transmitter/receiver systems, component failures, and so on.

2) *Convolutional changes.* These changes are the result of linear frequency alterations produced by the recording system, transmission channel, and the acoustic environment, such as room reverberation or high-frequency rolloffs.

3) *System noise.* Any noise contributed by the recorder and transmission systems, such as 60-Hz hum and high-level wow and flutter.

4) *Environmental noise.* Any unwanted noise added to the voice signal before it is sensed by the microphones, such as music, television broadcasts, ventilation sounds, and microphone-handling events.

5) *Large amplitude differences between talkers.* A problem often encountered with improperly recorded telephone conversations or when some individuals are much closer to a microphone than others.

6) *Signal losses.* The complete or partial loss of the audio information due to factors such as intermittent electronic components or cabling, damaged oxide surfaces on the magnetic tape, and low-power transmitters too far from the receiver.

The next step is an overall FFT review is usually conducted using a fairly fast exponential average that permits the instantaneous nonvoice signals to be somewhat minimized and the speech information to be displayed in a relatively real-time fashion. As an example, Fig. 7(A) shows a spectral average with no voice information and Fig. 7(B) the same recording with speech information present. A number of frequency characteristics of the recording are analyzed including the following.

1) *Speech frequency range.* Since enhancement of the voice information is the goal of the examination, a determination is made of the lowest and highest frequencies where speech signals are present. Usually different frequency bands on the FFT are utilized for the highest resolution.

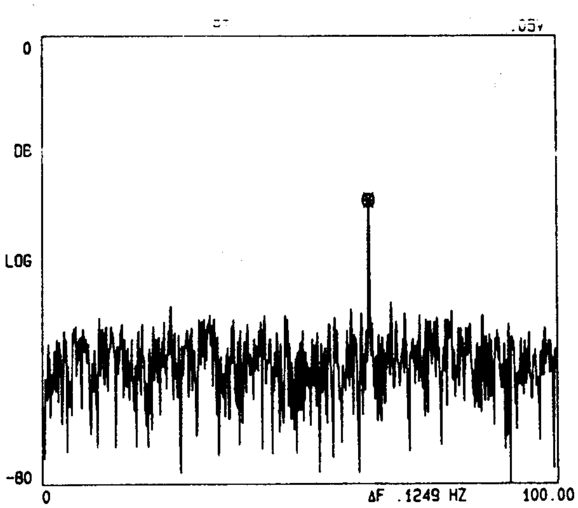


Fig. 6. FFT display of 63.5-Hz discrete tone that should play back at 60.0 Hz (0–100-Hz range).

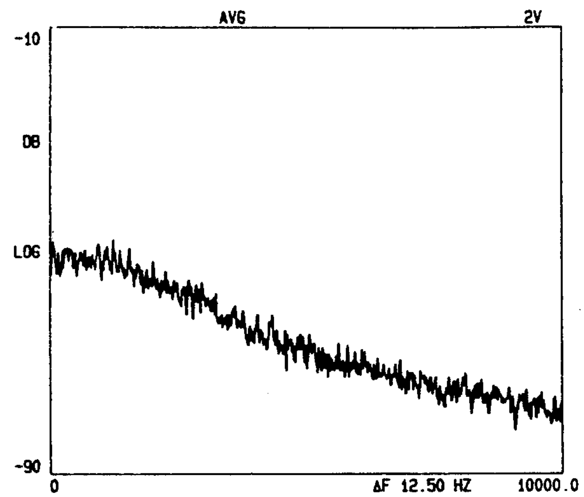
2) *Peak speech-to-noise ratio.* A rough determination is made of the level of the voice information compared with the background noise.

3) *Discrete tones.* The frequency and general amplitude of all high-level tones and their stability are noted.

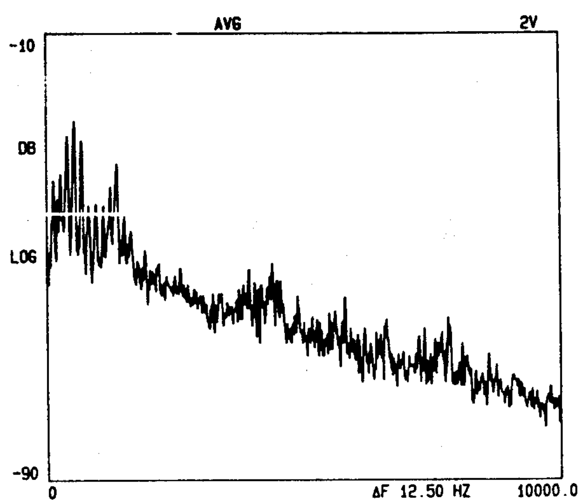
4) *Banded noise.* The general position, amplitude, and stability of fairly wide bands of noise are determined.

5) *Convolutional effects.* Using longer averaging modes, a determination is made whether the speech information is consistent with known long-term spectral responses, that is, generally a smooth amplitude rise peaking between 500 and 1000 Hz and then a slow rolloff out to beyond 10 kHz. Fig. 8(a) shows a normal long-term frequency average for a male speaker and Fig. 8(b) a highly convoluted one from a forensic recording of a room conversation.

Next, using the enhancement devices, the intelligi-



(a)



(b)

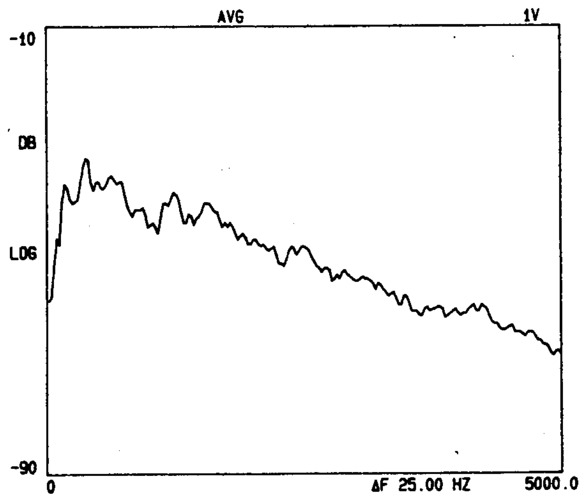
Fig. 7. FFT display of background sounds on a forensic recording (0–10-kHz range). (a) Without voice information. (b) With voice information.

bility of the speech information is improved by the attenuation of distortion, noise, and convolutional effects. However, the "golden rule" of enhancement is that no audio signals are removed or attenuated that decrease speech intelligibility, even slightly. If a recording sounds better overall by the reduction of a particular masking noise, but the understandability is somewhat reduced in the process, the noise is left in or lesser attenuation corrections are tried.

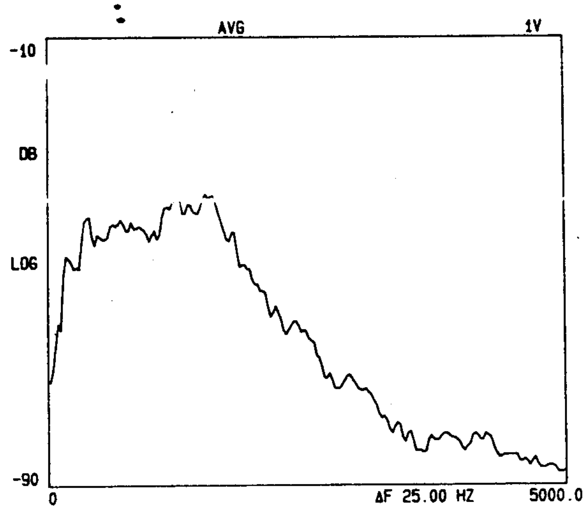
Some of the general procedures followed by the FBI in configuring the enhancement devices are detailed below, but the exact steps vary considerably among the different examiners.

1) Using sharply sloped bandpass filters, audio information above and below the speech frequency range is attenuated with both analyzer and aural verification.

2) If the voice spectrum is convoluted by, for example, considerable room reverberation, a digital deconvolution device or parametric equalizer is used to smooth and shape the frequency information. The equalizer usually functions better with distorted audio and very limited signal-to-noise ratios (SNR), since true wideband noise and nonlinear distortions are usually not time correlated.



(a)



(b)

Fig. 8. FFT display of long-term spectrum (0–5-kHz range). (a) Particular male speaker. (b) Particular male speaker on a typical forensic recording.

3) If sinusoidal tones are present in the recording from ac power leakage, fairly slow moving music, video arcade sounds, and other sources, the digital deconvolution devices are normally utilized, except again with distortion or limited SNRs, where notch, comb, or parametric filters are employed.

4) If a "buzz" sound is present, which is seen as periodic impulses in the time waveform, then comb filters are tried.

5) Banded noises from sources such as air conditioners, generators, or aircraft operation are usually handled by manual adjustment of the bandwidth and attenuation/boost controls on the parametric equalizers.

6) When one or more individuals in a conversation are recorded appreciably louder than other talkers or there are high-amplitude transientlike sounds present, then a gain-reduction device is used, usually after the recording has been processed through all the filters. Attack and release times are adjusted for natural-sounding speech and no loss of voice information, even when more than one individual is talking at the same time.

An aural check is made by the examiner at the end of the enhancement process to ensure that intelligibility has been improved. This is usually done by a direct A/B comparison of the original versus the enhanced audio. Stereo channels are processed separately to obtain the best intelligibility for each recording. After proper playback and cabling through the appropriate enhancement devices, copies are made of the improved audio signal for the contributor. The copies are normally prepared on standard cassettes with type I equalization and bias and/or open reels. To allow the most universal reproduction, monaural recordings are usually copied onto open reels of ¼-in-wide tape in a full or ½-track configuration at 3¾ in/s or standard cassettes of C90 length or shorter in a ¼-track configuration (both channels) at 1⅞ in/s. Stereo recordings are usually produced in a ½-track stereo configuration on the reel format and obviously ¼-track stereo on the cassette. The enhanced copies are marked with the laboratory number, the examiner's initials, and a description of the recording, such as "Enhanced copy of Q34," with a permanent-ink pen, and the safety tabs are removed on cassette copies. As a safeguard, a protection copy of the unfiltered audio is usually made and retained by the FBI in a secure storage area.

Work notes are taken of most facets of the enhancement process. This can include a description of the physical evidence, track configuration, azimuth error, speed error, aural and frequency observations, approximate length of the recording, all enhancement devices used with charts of setting when necessary, recorder and magnetic tape types, overall examination results, and any instructions that should be passed on to the contributor. A formal report is then sent to the contributor at the end of the examination, listing information such as FBI file and laboratory numbers, investigative title, evidence description and receipt date, results of examination, examiner's name, and disposition of the original evidence recordings and enhanced copies.

4 ENHANCEMENT EXAMPLES

To give some feel as to how an actual examination is conducted, two examples will be detailed. The first is a room conversation recorded on a standard cassette with a concealed microphone by a state police department and the second is a cockpit voice recording recovered in a major airplane crash.

The cassette is received via registered mail in a sealed package with a letter requesting that the clarity of the tape be improved and two enhanced copies prepared. After assignment of the laboratory and specimen numbers, the examiner marks and initials the housing, removes the still-intact safety tabs, and describes the evidence in the work notes. The tape looks to be in good physical condition, so it is briefly played back on a professional deck, a high-level area located, a short portion of the tape pulled from the housing, and the track developed with ferrofluid. The tape is found to be a ¼-track monaural configuration, left channel only, with apparently good azimuth alignment. The cassette tape is cleaned to remove the ferrofluid residue and rewound into its housing. To double-check the microscopic results, the frequency response is checked with the FFT analyzer on a deck with adjustable azimuth alignment and found to be consistent with playback using a properly aligned reproduce head. Playing back the recording on a ¼-track stereo deck with the left channel output, the 60-Hz tone power leakage signal is located at 61.5 Hz, so the variable speed is adjusted, while watching the analyzer display, to slow the transport to the correct playback speed.

Aural review of the tape reveals a poor-quality recording with room reverberation, limited high-frequency response, heating/air conditioner fan noise, and a range in voice amplitude from loud to very soft. The FFT reflects a convoluted spectrum in Fig. 9 and a fairly sharp rolloff above 500 Hz, though voice information is seen out to 4.8 kHz. Some of the spectrum is composed of banded noise from the ventilation system. The tape deck's output is cabled through a deconvolutional filter, which greatly attenuates the reverberation and flattens the frequency response, as shown in Fig. 10. This produces an increased voice-to-noise ratio and a more understandable recording since the higher speech frequencies, where voice intelligibility is greatest, have been boosted. The signal is next run through a parametric filter, which is set appropriately for the banded noise from the fan that was not completely removed by the previous filter. The procedure used is to identify the center frequencies and bandwidths from the analyzer, set the controls the same on the parametric equalizer, and then slowly attenuate each area until the maximum noise is removed without affecting speech intelligibility. Finally, the signal is cabled through two identical gain-reduction devices at 20:1 compression so that when the soft voices appear, they produce no amplitude reduction, but when the loudest voices are recorded, a drop of 10 dB per device is obtained. The use of two gain-reduction devices in series and moderate attack and release times often produces the best forensic results. The enhanced audio is then recorded on both channels of two professional stereo cassette decks using standard bias tapes. Detailed notes are taken of all the observations and procedures, and a final report, the evidence tape, and the enhanced copies are forwarded to the contributor.

In the second example, an aircraft cockpit recording tape is received from the U.S. Department of Justice requesting that the cockpit area microphone (CAM) channel be enhanced and three open-reel copies be prepared. The tape, which is ¼-in-wide reel tape, has been previously removed from its protective shell, leader tape added to each end, and placed on a 5-in open reel by personnel of the National Transportation Safety Board. Since these recordings are produced on endless-loop units, only the last 30 or so minutes before the crash will be present. Again, after receiving the laboratory and specimen numbers, the reel and the tape backing are marked and initialed with a permanent-ink pen. The tape and reel are found to be in excellent condition, so the tape is developed and a four-in-line track configuration is identified. Azimuth alignment looks slightly off under the microscope, so after cleaning the tape where the ferrofluid was used, the tape is placed on a tape recorder with an adjustable head stack (Fig. 2). Playback reveals that the 400-Hz power leakage frequency is at 412.4 Hz, so the variable-speed control is used to correct the error. By adjusting the reproduce head up and down, the designated channel is located and the azimuth adjusted so the output is maximized for both amplitude and frequency response, with no

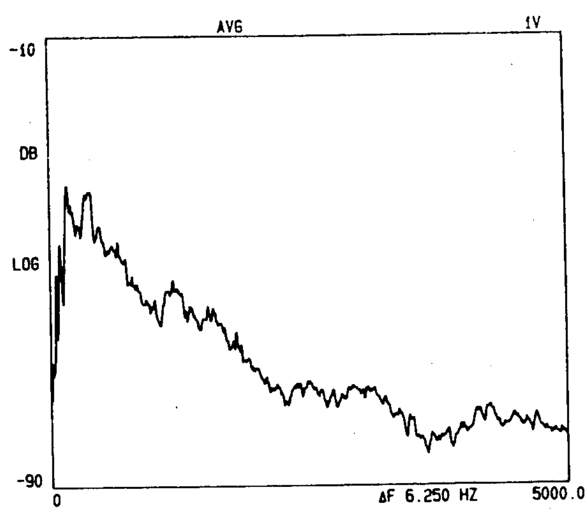


Fig. 9. FFT display of room recording before processing (0–5-kHz range).

crosstalk from other channels, using the FFT analyzer. The other three channels on the tape are found to contain good-quality radio transmissions.

Aural review reflects that the voice information is at or below the random noise present in the cockpit and discrete tones can be heard. FFT analysis reveals a high-level 400-Hz signal, a somewhat lower 800-Hz discrete tone, and voice information up to at least 3 kHz. The tape reproducer's output is cabled through sharply sloped bandpass filters at 200 Hz for the high pass and 4 kHz for the low pass. Further reduction of the passband had a slightly negative effect on the voice information. Two notch filters are then used to attenuate the 400- and 800-Hz tones, but only to the point of inaudibility. If deeper notches than necessary are used, the recording can sound unnatural. Attempts at filtering with other devices are not helpful, so the enhanced signal is recorded simultaneously on three open-reel recorders at 7½ in/s in a full-track configuration on 7-in reel with standard bias tape. As before, appropriate notes are taken, a final report is prepared, and the evidence and copies are returned to the contributor.

5 TRAINING

The Engineering Section of the FBI follows a rigorous procedure in qualifying individuals as examiners in the field of forensic tape enhancement. This includes screening of potential applicants, a lengthy apprenticeship, attendance at specialized schools, moot court training, and formal approval by senior examiners and supervisory personnel. Even after certification, training is continued through supervisory reviews, additional schooling, and regular hearing tests.

The evaluation of applicants includes affirmation of a Bachelor of Science or higher degree from an accredited college or university, contact with past employers and work associates, and verification of excellent hearing. In addition, at least two formal interviews of the applicant are conducted and a Top Secret background clearance is performed, since some of the recordings enhanced by the FBI are classified. Once hired, or reassigned from other FBI duties, the trainee examiner is placed in an apprenticeship program under the direction of a fully qualified examiner, which lasts for at least one year for individuals with an appropriate science degree and three years or more for others. This training period includes full-time experience on over 300 submitted recordings using laboratory tape recorders, analyzers, filters, and so on. Though the technical procedures are emphasized, other important areas such as proper note taking, report preparation, and evidence handling are covered in detail. Concurrently, the trainee attends lectures, demonstrations, schools, workshops, and conventions concerning various laboratory devices, recording theory and practice, audio engineering topics, and related subjects.

When the training examiner is completely satisfied that the individual has a good mastery of tape enhancement techniques and equipment, moot court training is then given to assess verbal responses and demeanor under courtroom conditions. The moot court exercises are made as stressful as possible with experienced FBI personnel and lawyers acting as judges, attorneys, and jury members. Wide latitude is allowed in the questioning to force the nearly trained individual to cope with difficult legal and technical concepts that are often encountered during cross-examination. With the concurrence of supervisory personnel and other examiners, the trainee is approved to receive and conduct enhancement cases with an overview process continuing for a few months.

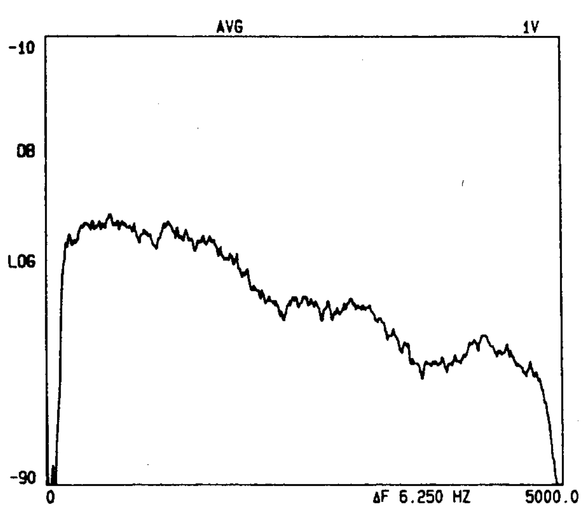


Fig. 10. FFT display of room recording after processing by deconvolutional filter (0–5 kHz range).

6 EVIDENCE HANDLING

Since enhancement examinations utilize original tape recordings that may subsequently be admitted in criminal and civil proceedings, evidence handling procedures are exceedingly important. Every step from receipt until return to the contributor is carefully monitored and documented. Areas of concern include custody, storage, and transport of the original recordings and the enhanced copies.

The chain of custody of the evidence from date of receipt until its release to the contributor or other party is set forth in written form—ledgers, work notes, signed receipts, evidence envelopes, and so on. While in the laboratory, the evidence is usually assigned to only one examiner to simplify accountability and possible future testimony. Evidence is stored at normal room temperature in a security file or a sturdy lockable cabinet in an area that is secured when unattended. In addition the laboratory is housed in a secured building. Access to the evidence is strictly limited to necessary personnel and is stored well away from magnetic fields produced by loudspeakers, transformers, and other devices. Transport of evidential recordings is only handled in four ways: 1) registered mail, 2) overnight delivery services with signature confirmation (no express mail),

3) courier, and 4) personal delivery. All, except personal delivery, are forwarded in a sealed condition with at least 3 in of packing between the tapes and the outside of the box to avoid the possible effects of stray magnetic fields and improper handling.

7 TESTIMONY

Even after a recording has been enhanced successfully, the improved copies sometimes cannot be used at trial without expert testimony. To allow accurate and meaningful testimony, the FBI examiners prepare a qualification list, attend a pretestimony conference with the attorney, present a proper appearance and demeanor on the stand, and verbalize the important aspects of the examination and chain of custody in an understandable way to the judge and jury.

A qualification list allows the presenting attorney to ask the appropriate questions of the examiner to reflect his or her training, education, professional societies, and so on. The list often includes the following, as appropriate:

1) Present title, organization, responsibilities, and length of service

2) Pertinent prior employment information

3) Formal college and university degrees, and additional college courses

4) Pertinent technical schools, seminars, etc.

5) Membership in professional societies

6) Publications in the tape analysis field

7) Number of times previously qualified as an expert in court

8) Approximate number of different investigative matters in which examinations have been conducted

9) Approximate number of different evidential recordings analyzed.

Discussion of the examination with the attorney before testimony is important to explain laboratory methodology, set forth results, provide the qualification list, and determine the specific questions that will be asked of the examiner on the stand. The attorney can provide guidance on local court procedures, the expected questions from opposing counsel, and exactly when the testimony will be needed. Normally the attorney is advised that for clarification, two particular questions should be asked near the end of the examiner's testimony: "Did the enhancement process in any way change the original evidence tape?" and "On the enhanced copies, did the process in any way change what the talkers actually said?" The answers to both questions are obviously no, and an explanation is then given to the court.

Examiners dress in proper business attire and direct explanations to the jury, when present, to allow feedback on their understanding of the answers. They are trained to maintain a proper demeanor under the stress and distractions of, for example, yelling opposing counsel, interruptions by court reporters, and inattentive jury members. Although the FBI does not conduct examinations for the defense in criminal matters, the examiners will appear, when requested, for either side at trial to testify to the results of the analyses, with all salary, examination, and travel expenses paid by the FBI.

8 SUMMARY

With careful selection and training of examiners, high-quality enhancement devices, proper evidence handling, and effective testimony, the FBI has formalized a procedure to enhance tape recordings produced by law enforcement and other agencies involved in forensic investigations.

9 ACKNOWLEDGMENT

The author would like to thank the following FBI associates who have contributed to this report and to the improvement of tape enhancement in general: Steven A. Killion, Donald V. Ritenour, Jr., James B. Reames, Mark S. Britton, George W. Clark, Jr., James P. Foye (now with the U.S. Postal Service), Jeffrey A. Granger, Robert L. Halvorson, Noel Herold, Shawn M. Hoffman, Artese S. Kelly, Barbara A. Kohus, Dale G. Linden, John L. Losinski, Gregory J. Major, Kenneth A. Pfarr (now with the U.S. Secret Service), James J. Ryan, Jr., Keith Sponholtz, Richard L. Todd, and Dennis O. Williams. Special thanks go to James E. Paul, Digital Audio Corporation, Raleigh, North Carolina, for his assistance regarding deconvolutional filters.

10 REFERENCES

[1] N. F. Ramsey, L. W. Alvarez, H. Chernoff, R. H. Dicke, J. I. Elkind, J. C. Feggeler, R. L. Garwin, P. Horowitz, A. Johnson, R. A. Phinney, C. Rader, and F. W. Sarles, "Report of the Committee on Ballistic Acoustics," Commission on Physical Sciences, Mathematics, and Resources, National Research Council, National Academy Press, Washington, DC, 1982.

[2] B. E. Koenig, "Acoustic Gunshot Analysis—The Kennedy Assassination and Beyond," *FBI Law Enforcement Bull.*, vol. 52, no. 11, pp. 1–9; no. 12, pp. 1–9, 1983. Reprinted in *Internat. Criminal Police Rev.*, vol. 40, no. 384, pp. 2–11; no. 385, pp. 43–53, 1985.

[3] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1975), pp. 356–437.

[4] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1975), pp. 284–320.

[5] S. D. Stearns, *Digital Signal Analysis* (Hayden, Rochelle Park, NJ, 1983), pp. 74–88.

[6] D. G. Fink, Ed., *Standard Handbook for Electrical Engineers*, 11th ed. (McGraw-Hill, New York, 1978), pp. 2–56.

[7] B. E. Koenig, "Spectrographic Voice Identification: A Forensic Survey," *J. Acoust. Soc. Am.*, vol. 79, p. 2089 (1986 June).

[8] J. E. Paul, "Adaptive Digital Techniques for Audio Noise Cancellation," *IEEE Circuits Sys.*, vol. 1, pp. 2–7 (1979 Dec.).

[9] B. E. Koenig, "Tape Recorder Azimuth Misalignment," *Crime Lab. Dig.*, vol. 15, pp. 44–54 (1988 Apr.).

[10] B. E. Koenig and B. A. Kohus, "Measurement of Recorder Speed Changes in Authenticity Examinations," *Crime Lab. Dig.*, vol. 14, pp. 139–152 (1987 Oct.).

THE AUTHOR

Bruce E. Koenig received a B.S. degree in physics and mathematics from the University of Maryland in 1968, an M.F.S. degree in forensic science from George Washington University in 1977, and has taken additional courses at George Mason University, Massachusetts Institute of Technology, and the University of Utah. In 1970 he became a special agent (SA) of the Federal Bureau of Investigation (FBI) and served in the Atlanta and Detroit divisions investigating bank robberies, federal prison escapes, and other violations of federal law. In 1974 he became a supervisory SA in the engineering section at FBI headquarters and has since been involved with the analysis of magnetic tape recordings produced or collected by federal organizations or law enforcement agencies worldwide. These have included tapes containing undercover drug buys, political corruption, bribes, racketeering, and "black box" cockpit conversations of major airlines' disasters.

Mr. Koenig's responsibilities include the analysis of audio and video tapes to improve intelligibility, comparison of voice samples, identification of nonvoice signals, and authentication of recordings. He has examined over 7000 separate tapes for law enforcement agencies in all 50 states, the territories, and eighteen foreign countries, and has qualified in court as an expert witness on over 170 occasions.

Mr. Koenig is a member of the Audio Engineering Society, Acoustical Society of America, and the Institute of Electrical and Electronics Engineers and has published articles on forensic tape analysis in publications including *the Journal of the Acoustical Society of America*, *Crime Laboratory Digest*, *FBI Law Enforcement Bulletin*, and *International Criminal Police Review.*

Speech production and syntax comprehension deficits on Mt. Everest

Philip Lieberman, M.S., Ph.D. and Athanassios Protopapas, B.S., M.S.
Department of Cognitive & Linguistic Sciences
Brown University
Barbara G. Kanki, M.S., Ph.D.
NASA-Ames Research Center
Moffet Field, CA

Aviation, Space, & Environmental Medicine, in press

Deficits in speech motor control and in the comprehension of syntax were observed as five members of the 1993 American Sagarmatha Expedition ascended Mt. Everest. We analyzed speech recordings and cognitive test scores of the climbers at different altitudes. The mean "voice onset time" interval that differentiates "voiced" stop consonants from their "unvoiced" counterparts (e.g., a [b] from a [p]) decreased from 24.0 ms at Base Camp to 5.4 ms at Camp Three. The time needed to comprehend simple spoken English sentences increased by 50% at higher altitudes, and was correlated with speech motor deterioration. This pattern of deficits is similar to that noted for Parkinson's disease and may reflect disruption of subcortical pathways to prefrontal cortex. Similar procedures could be used to remotely assess cognitive impairments caused by hypoxia, carbon monoxide or alcohol intoxication, or drugs, in order to monitor crew behavior in aeronautics and spaceflight operations, or to evaluate the treatment of neurodegenerative diseases such as Parkinson's disease.

Keywords: Hypoxia, Cognitive deficits, Mountaineering, Speech production measurements, Language tests

In the seventy odd years since the ascent of Everest was first attempted climbers have frequently reported motoric and cognitive deficits at extreme altitude. Impaired judgement is implicated in many of the fatalities occurring on Everest and other 8,000 m mountains (Ward, Milledge, & West, 1989; Nelson, Dunlosky, White, Steinberg, Townes, & Anderson, 1990). Several researchers have investigated the effects of high altitude on cognitive and motor performance. Most studies have been concerned with lasting effects after exposure to high altitude; few studies have tested mountaineers at high altitudes to investigate possible transient effects. Apart from the theoretical and scientific interest in the cognitive and motoric deficits at high altitude, such research may lead to a practical, online, unobtrusive system for remote monitoring of the effects of situation-specific impairments, such as high altitude climbing or flying, in order to reduce the risks associated with these activities.

The research was conducted under consortium agreement NCC2-5041 from NASA Ames Research Center to Brown University. It was supported by NASA, Office of Life and Microgravity Sciences and Applications. We thank the members of the 1993 American Sagarmatha Expedition for participating in the research, Eric Reed and J. William Youngs for their essential role in data acquisition at Everest Base Camp, and Emily R. Pickett for comments on the manuscript. We would also like to thank the Motorola Corporation and the Sony Corporation of America for the equipment support necessary to carry out this research on Mt. Everest.

Temporary impairments in cognitive functioning found at high altitude include deterioration of the ability to learn, remember and express information verbally (Townes, Hornbein, Schoene, Sarnquist, & Grant, 1984), impaired concentration and cognitive flexibility (Regard, Oelz, Brugger, & Landis, 1989), decline of the feeling of knowing (Nelson et al., 1990), and mild impairment in either short-term memory or conceptual tasks (Regard, Landis, Casey, Maggiorini, Bärtsch, & Oelz, 1991). Kennedy, Dunlap, Banderet, Smith, & Houston (1989) reported impairments in grammatical reasoning and in pattern comparison during a slow, multi-day, simulated ascent in a hypobaric chamber. Cognitive deficits found in climbers after a high-altitude expedition include decreased memory performance (Cavaletti, Moroni, Garavaglia, & Tredici, 1987), mild impairment in concentration, verbal learning and memory, and cognitive flexibility (Regard et al., 1989; Oelz, Regard, Wichmann, Valavanis, Witztum, Brugger, Cerretelli, & Landis, 1990), and decline in visual and verbal learning and memory (Hornbein, Townes, Schoene, Sutton, & Houston, 1989). It is unclear whether any of these deficits become permanent after repeated prolonged exposure to extreme altitude, as findings from different studies have reached opposing conclusions (cf. Clark, Heaton, & Wiens, 1983; Jason, Pajurkova, & Lee, 1989).

A possible explanation for the reported deficits is that brain hypoxia, caused by lowered oxygen content of the inspired air, selectively impairs some brain structures. Histologic studies of the hypoxic brain have identified regions of "selective vulnerability" to hypoxia in the hippocampus, cerebellum, layers III, V, and VI of the neocortex, and the basal ganglia (Brierley, 1976). According to current theories of brain function, the compromised cells in the hippocampus and the cerebellum may account for some learning (memory) and motor deficits, respectively; the consequences of the affected cortical regions may be quite extensive, possibly including deteriorations of perception, planning, and evaluation of danger. However, the possible effects of basal ganglia dysfunction at high altitude have not been investigated. In order to assess these effects, we need to take into account studies that relate lesions in the basal ganglia to particular types of deficits.

Recent studies of Broca's aphasia (Baum, 1988; Baum, Blumstein, Naeser, & Palumbo, 1990) and Parkinson's disease (Grossman, Carvell, Stern, Vernon, & Hurtig, 1991; Lieberman, Kako, Friedman, Tajchman, Feldman, & Jiminez, 1992) show deterioration of speech motor control, deficits in syntax comprehension, and other cognitive deficits. Such decrements may reflect the degradation of subcortical basal ganglia pathways to prefrontal cortex (Metter, Kempler, Jackson, Hanson, Mazziota, & Phelps, 1989; Metter, Riege, Hanson, Phelps, & Kuhl, 1984; Lieberman, 1991; Lange, Robbins, Marsden, James, Owen, & Paul, 1992; Cummings, 1993). It is now known that many pathways involved in motor control as well as in higher associative or cognitive functions include connections to and from the basal ganglia (Parent, 1986). These pathways may also be implicated in the motoric and cognitive deficits reported by climbers at extreme altitude. In this study we administered a battery of tests, for which we had comparative data from Parkinson's disease, to climbers at extreme altitude. We measured speech motor control, comprehension of meaning conveyed by syntax, and "frontal" cognitive functions. These tests can be administered remotely with minimal equipment.

The speech attribute that we studied, Voice Onset Timing (VOT), differentiates English "voiced stop" consonants like [b], [d], and [g] from their unvoiced counterparts [p], [t], and [k], respectively. In order to produce a [b], a speaker has to initiate "phonation" (i.e, quasi-periodic vibration of the vocal folds) soon after opening the lips (within about 20 ms) to release the pressure built in the vocal tract. In contrast, phonation is delayed for 40 ms or more after lip opening in a [p]. Similar timing distinctions differentiate [d]s from [t]s and [g]s from [k]s. Figure 1 shows the waveforms for a [b] and a [p] produced by the same speaker, where the lip opening (identified by a visible burst) and the onset of phonation (evidenced by periodicity in the waveform) have been marked. The time delay between the marks is the VOT. Normally, speakers of English and many other languages maintain the VOT distinction between voiced and unvoiced word-initial stop consonants by keeping the VOT regions of the two separated by at least 20 ms. Listeners make use of this cue to differentiate stop consonants in word-initial position (Lisker & Abramson, 1964).

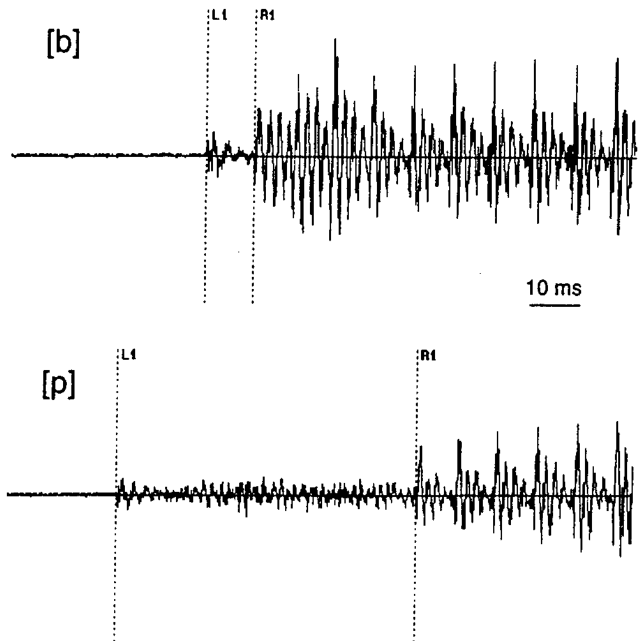


Figure 1. Speech waveform segments corresponding to a [b] and a [p] spoken by the same speaker under identical conditions. Cursors have been placed at the onset of the burst that was caused by opening the lips (L1) and at the onset of periodicity that indicates vocal fold vibration (R1). The marked interval, Voice Onset Timing (VOT) is used by speakers and listeners to differentiate the two types of consonants in word-initial positions.

Syntax comprehension was assessed by the Rhode Island Test of Language Structure (RITLS), a test initially designed to evaluate hearing impaired children. The RITLS assesses the extent to which subjects are able to use syntactic properties of sentences (word order, markers of the relationships between clauses, and markers of non-canonical order) to understand them. It includes "simple" sentences (consisting of a single clause) and "complex" sentences (containing embedded clauses), presenting a representative sample of the syntactic structures of English. Vocabulary and morphology are tightly controlled and sentence length is balanced between simple and complex sentences. The vocabulary is kept very simple and none of the sentences are very difficult; normal 10 year old native English speaking children make almost no errors on either the simple or the complex sentences (Engen & Engen, 1983). In this study we measured processing difficulty by timing the subjects' responses.

Methods

Subjects

The subjects were five male members of the 1993 American Sagarmatha Expedition team to Mount Everest. Their ages ranged from 35 to 52 years and their education from High School to M.D. Informed consent was obtained from all members of the original team (eight males and one female); however, some members dropped out of the study for a variety of reasons. All the subjects were experienced climbers (each had at least eight years of experience on major peaks). Sherpas worked closely with the team but they did not officially participate in the study. The study procedures were approved by the Brown University Human Subjects Committee.

Materials and Locations

The expedition approached Mount Everest by the "normal" South Col route, which involves establishing a series of high camps over a period of two months. We gathered data from the climbers at Base Camp (altitude 5,300 m) before and after the ascent to the higher camps, at Camp Two (6,300 m), and at Camp Three (7,150 m). Conditions at Camp Four (8,000 m) or higher (e.g., weather conditions and communications quality) made testing problematic, particularly voice recordings. No data from locations higher than Camp Three are reported.

Supplementary oxygen was not used below Camp Four. All testing sessions were held within a day of reaching each location. Final testing at Base Camp was done after a descent from the summit for four subjects; one subject who decided to forgo a summit attempt was tested after his descent from Camp Four. The experimenters remained at Base Camp throughout the expedition, where they recorded speech samples transmitted from the climbers. Motorola hand-held Sabre radios (5W RF output), Max-Trax base station radios (20W RF output), and Sony NT-1 digital micro DAT recorders were used throughout the project. Sampling with the NT-1 DAT recorder was done at 27 KHz using 12-bit logarithmic quantization equivalent to 16-bit linear. The signals were resampled for computer analysis at 20 KHz using 12-bit linear quantization. The digital recording established an accurate time base for the VOT measurements.

Speech measurements

Speech samples for the VOT analysis were obtained by asking each subject to read 60 English monosyllabic words (a 30 word list read twice) that had voiced and unvoiced stop consonants in initial position and final position, e.g., *bat, kid,* etc. VOT was subsequently measured at Brown University using an interactive computer-implemented system. Cursors were placed on the onset of the burst produced on the release of each word-initial stop consonant and at the onset of phonation, by means of both visual inspection of the waveform and by listening to marked portions of the signal.

VOT measurements from all three places of articulation (i.e., labial [b] and [p], alveolar [d] and [t], and velar [g] and [k]) were combined by aligning their perceptual boundaries.[1] The *separation width*, i.e., the distance (in time units) between the longest voiced VOT and the shortest unvoiced VOT, was measured for each subject at each location. Deterioration in motor control is manifested by reduced separation width; in cases of severe impairment the voiced and unvoiced regions might overlap and the separation width would become negative.

Syntax testing

A 50 sentence version of the RITLS was administered at each location. Each version included 25 simple and 25 complex sentences balanced for vocabulary, sentence length, and syntactic patterns. RITLS test booklets containing the sketches corresponding to the sentences were carried to the higher Camps. The test was administered by showing the subject a page which presented three elaborated line drawings, one of which best exemplified the meaning of the sentence that was then read aloud by the experimenter. For example, for the sentence "The man is watching the girl who is in the water" the choices were (1), a man and a girl on the sand, (2), a man on the sand and a girl in the water, and (3), a man in the water and a girl on the sand. The subject then responded by announcing the number of the sketch that best exemplified the meaning of the sentence.

Before each sentence the subject announced the page number he was looking at to indicate that he was ready and to verify that he was looking at the correct drawings. The test sentences, which were read aloud by the experimenter, and the subject's vocal responses were tape recorded. The response time was determined by a single listener by measuring with an electronic stopwatch the time interval between the end of each spoken sentence and the subject's response. Multiple measurement of several sample trials showed that such measurements were consistent within 0.1 s.

Cognitive testing

Three cognitive tests were also administered to subjects at each location, to test attention and concentration, expressive language and structured response initiation, and maintenance and shifting of cognitive sets (Parkinson's study group, 1989). The confrontation "naming" test, sometimes referred to as the verbal fluency test, tested the subjects' ability to

[1] For a given set of VOT measurements and a "boundary" point on the VOT axis, the corresponding "perceptual overlap" is the percentage of VOT measurements on the "wrong" side of the boundary, with a 10 ms tolerance, i.e., the sum of all voiced VOTs greater than 10 ms less than the assumed boundary plus the sum of all unvoiced VOTs less than 10 ms greater than the assumed boundary divided over the total number of VOT measurements. The perceptual boundary is defined as the boundary point for which perceptual overlap is minimized; if there is a region of minimum overlap the perceptual boundary is set at its midpoint.

generate words beginning with particular letters of the alphabet. The subject was presented with a letter and was asked to produce as many words as possible in one minute, excluding proper nouns; this was repeated with two more letters, for a total of three letters at each location. Different letters were used at each location. The subject's score was the total number of words produced for all three letters.

The digit-span tests tested the subjects' ability to repeat a sequence of numbers in the order presented (forward digit span) and in the reverse order (backward digit span). To administer this test the experimenter read aloud a sequence of digits, starting with a sequence of three digits. The subject had to immediately repeat the digits in the same order; this was repeated once more with another sequence of the same length. If the subject correctly repeated at least one of the two sequences the process was continued with two sequences of length greater by one. When the subject failed to repeat both sequences of a given length the test was terminated and the subject's score was the total number of sequences reproduced correctly. Then the same process was repeated (with different sequences, starting with a sequence two digits long) but this time the subject was required to reproduce the sequence in reverse order.

Finally, the odd-man-out test tested the subjects' ability to form and then shift abstract categories. A test booklet, carried up to Camps Two and Three, contained on each page a set of three figures, one of which shared a feature with each of the other two. For example, there was a page with a large oval, a small oval, and a large triangle. The subject's task was to form a criterion (e.g., either size or shape) and to pick the "odd" figure on each page according to that criterion. After this first sort the subject was asked to sort the same set of figures using another criterion. Subjects were not told what the possible criteria were. The total number of errors (in both sortings) for each subject were counted.

In addition to these tests, the expedition maintained a log book noting each day's activities. The experimenters noted any incidents that appeared to exemplify poor judgement on the part of the climbers.

Table 1
Separation width (in milliseconds) between longest voiced VOT and shortest unvoiced VOT for each subject at each location with the three places of articulation combined (BB=Base Camp before the climb, C2=Camp Two, C3=Camp Three, BA=Base Camp after the climb).

	Location			
Subject	BB	C2	C3	BA
1	30	26	13	29
2	22	14	6	7
3	32	8	8	20
4	14	13	-10	16
5	22	-3	10	23

Results

Table 1 shows the separation width of each subject's VOTs on each location for all places of articulation combined. Note that there is a wide VOT separation at base camp, but the distinction becomes less pronounced at the higher camps. This drop is illustrated in Figure 2, where the labial stop VOTs of Subject 1 occupy two distinct, well separated, regions at Base Camp before the climb (2a) but are less separated at Camp Two (2b) and even less so at Camp Three (2c). In some instances, separation decreased considerably at the higher camps, and even overlap occurred.

The subject averages are plotted in Figure 3. Analysis of variance showed a significant effect of location ($F(3,12)=6.30$, $p<0.008$). Pairwise contrasts showed that the differences between Base Camp before the climb (24.0 ms) and Camp Three (5.4 ms), and between Base Camp after the climb (19.0 ms) and Camp Three were significant ($F(1,4)=62.22$, $p<0.001$, and $F(1,4)=11.52$, $p=0.027$, respectively); the difference between Base Camp before the climb and Camp Two (11.6 ms) was marginally significant ($F(1,4)=5.99$, $p=0.071$). Note that the mean separation width at the higher camps is less than the normal 20 ms which are considered necessary for our perceptual system to unambiguously perceive the categorical distinctions between "voiced" and "unvoiced" stop consonants.

The mean response time (RT) to the RITLS items is shown in Table 2, separately for the simple and the complex sentences. Note that RT increases at the higher camps, indicating an increased difficulty in syntactic processing. Figure 4 plots the subject averages for four subjects[2] at all locations. Analysis of variance showed significant main effects of complexity ($F(1,3)=22.57$, $p=0.018$) and of location ($F(3,9)=5.11$, $p=0.025$) but no interaction ($F(3,9)=0.95$, $p=0.455$). In analyses of variance separately for simple and for complex sentences, there was a significant effect of location on simple sentence RTs ($F(3,9)=10.18$, $p=0.003$) but not on complex sentence RTs ($F(3,9)=0.95$, $p=0.458$). Pairwise contrasts using only RTs to simple sentences showed a significant difference between Base Camp before the climb (1.85 s) and Camp Two (2.40 s, $F(1,3)=44.34$, $p=0.007$) and between Base Camp before the climb and Camp Three (2.78 s, $F(1,3)=46.33$, $p=0.006$). The drop in response time after returning to Base Camp (i.e., between Camp Three and Base Camp after the climb) was marginally significant ($F(1,3)=7.63$, $p=0.070$).

The Pearson product-moment correlation coefficient between the subjects' response time to simple sentences and their VOT separation width was -0.774, (significant to $p=$

[2]Data for Subject 2 at Camp Three are missing because of a recording error. Using the mean value of the other four subjects' RTs at Camp Three gives similar results in all statistical tests reported here.

Table 2
Mean response time (in seconds) to the RITLS sentences by subject and location, separately for the simple and the complex items of the test.

	Base Camp, before		Camp Two		Camp Three		Base Camp, after	
Subject	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex
1	1.47	2.75	1.93	2.53	2.34	3.36	2.17	2.65
2	1.93	3.19	2.48	2.73	(*)	(*)	2.36	2.52
3	1.99	3.28	2.37	2.91	3.37	2.97	1.96	3.27
4	2.10	3.03	2.72	4.07	3.18	3.18	2.37	3.40
5	1.83	1.66	2.57	3.17	2.59	2.97	1.56	1.78

*These data are missing because of a recording error.

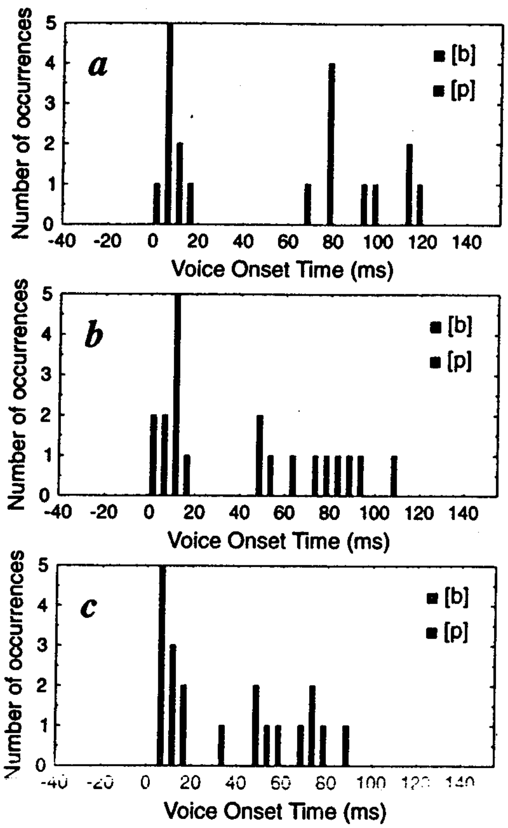


Figure 2. Number of VOT measurements per 5 ms bin. (a) VOTs for Subject 1's labial stop consonants ([b]s and [p]s) at Base Camp before the climb. Note that the [b]s cluster between 0 and 20 ms, whereas the [p]s occupy a distinct range after 70 msec. The separation width is about 52 ms, so the [b]s can be readily distinguished from the [p]s by virtue of the VOT distinction. (b) VOTs for Subject 1's labial stop consonants at Camp Two. Note that the separation width has decreased to 26 ms. (c) VOTs fot Subject 1's labial stop consonants at Camp Three. The separation width for these consonants decreased to 13 ms, thus (given a perceptual tolerance of about 20 ms) preventing absolute differentiation between [b] and [p] on the basis of VOT only.

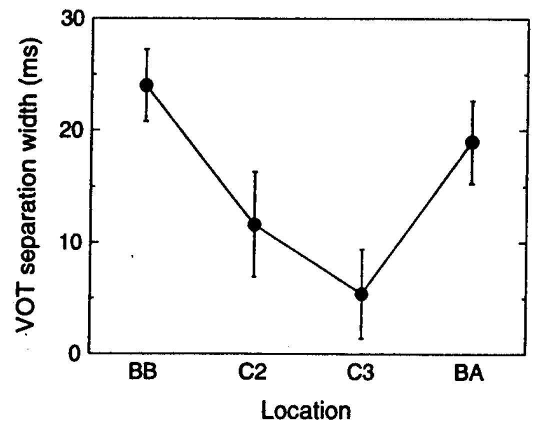


Figure 3. Mean VOT separation width for five subjects at four locations (BB=Base Camp before the climb, C2=Camp Two, C3=Camp Three, BA=Base Camp after the climb). Error bars show standard error.

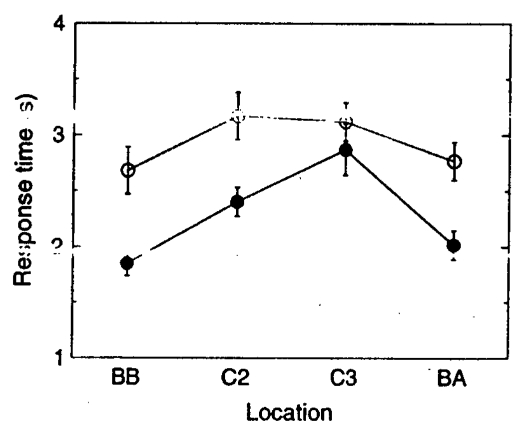


Figure 4. Mean response time to the simple (•) and complex (o) sentences of the RITLS for four subjects at four locations (BB=Base Camp before the climb, C2=Camp Two, C3=Camp Three, BA=Base Camp after the climb). Error bars show standard error.

Table 3
Total number of words generated for three initial letters in the confrontation naming test, five subjects at four locations (BB=Base Camp before the climb, C2=Camp Two, C3=Camp Three, BA=Base Camp after the climb).

	Location			
Subject	BB	C2	C3	BA
1	28	27	16	28
2	28	35	28	30
3	29	25	(*)	26
4	22	30	24	22
5	27	25	29	30

*These data are missing because of a recording error.

0.0001), which means that one can account for 60% of the syntax comprehension variance by the VOT measure.[3]

The total number of words generated by each subject in the confrontation naming task, shown in Table 3, was not significantly affected by testing location ($F(3,12)=1.14, p=0.37$). The individual subjects' performance on the digit span tests was unaltered throughout the experiment ($F(3,12)=1.06, p=0.40$); the subjects' scores are shown in Table 4. Almost no errors were made on the odd-man-out test except for Subject 4 after a respiratory infection on the mountain. His error rate was 30% on the second trial at Base Camp after he had returned from Camp Four; VOT overlap also occurred after this infection.

Several episodes occurred in which subjects' judgement was remarkably compromised. For example, Subject 4 (whose VOTs overlapped at Camp Three) advocated climbing in extreme avalanche conditions and was only dissuaded after vehement discussion. Another climber would have fallen into a wide crevasse that he was about to jump unroped had a Sherpa not intervened.

Discussion

We have found a significant effect of altitude on VOT separation width and on simple sentence comprehension response time. In agreement with Regard et al. (1991), who used similar tasks, no effect was found for the digit span, the confrontation naming, or the odd-man-out test. The observed deficits appeared to be temporary, since performance improved upon return to Base Camp. This finding is consistent with previous studies that have found no persistent impairments after a high altitude expedition (Clark et al., 1983; Jason et al., 1989). Hornbein et al. (1989) also reported intact performance in the digit span test after a simulated ascent to 8,848 m.

We conclude that hypoxia, caused by low concentration of oxygen in the air, caused subjects' neural functioning at high altitude to depart from normal, at least in the regions of the brain involved with syntax comprehension and speech motor control. Complex sentences were found to take more time than simple sentences to process, therefore reaction time can be used to assess processing difficulty.[4] The increase in response time to the simple sentences at higher altitudes, indicating greater processing difficulty, suggests that the climbers' neural functioning is considerably compromised. Note that 10 year old children have no problem understanding any of the RITLS sentences. In this light, it may be less surprising that climbers and pilots have often reported impaired judgement at high altitudes. The fact that the similar trend of response times to complex sentences is not significant may be due to the small number of subjects studied. Alternatively, it is possible that complex sentences, because of their embedded clauses, may make heavy demands on processing resources that are already impaired at Base Camp.

The small number of subjects tested may have affected the power of our statistical tests. However, apart from the effect of altitude on response time to the complex items of the RITLS, there is no clear trend in any of the other tests. Therefore, it is not very likely that the lack of statistical significance is a result of the small number of subjects; a genuine lack of effect of the factors tested seems more likely. This null result leads us to the conclusion that some tasks were unaffected at high altitudes. Practice effects, although probably playing some role in the subjects' performance at the higher camps, cannot alone account for the observed pattern of results. If subjects were improving at the experimental tasks and this improvement was offset by cognitive slowing down at the high camps, one would expect a small or no significant effect of altitude during the ascent, as observed, and a substantial improvement at Base Camp after the climb to the summit, which is not consistent with our findings. Alternatively, there may have been no such improvement upon return to Base Camp because of some longer lasting deficit caused at high altitude. Such a hypothesis is neither supported by others' findings with similar testing nor by the improvement in RITLS simple items response times upon return to Base Camp. Although one should be cautious when interpreting statistical tests with small N, particularly when referring to cognitive functions (notoriously variable between individuals), the overall pattern of our findings indicates that cognitive impairments when climbing to high altitude are selective and temporary.

Other factors that might have caused the observed deficits

[3]The data of Subject 2 for three locations were included in the correlation; leaving them out made no difference ($r=-0.776, p=0.0004$).

[4]Regarding reaction times in hypoxic situations, Fowler & Kelso (1992) reported a slowing of "the preprocessing state of stimulus evaluation." Their findings do not apply to our situation because they were for responses to visual stimuli, whereas earlier studies had found little effect of hypoxia on the latency of evoked potentials for auditory stimuli (Deecke, Goode, Whitehead, Johnson, & Bryce, 1973). Moreover, if the increased RTs in our tests were due to some general slowing of responding it should have been identical for simple and complex sentences.

Table 4
Digit span (total number of sequences correctly repeated) forward and backward for all five subjects at all locations.

	Base Camp, before		Camp Two		Camp Three		Base Camp, after	
Subject	Forward	Backward	Forward	Backward	Forward	Backward	Forward	Backward
1	10	5	9	2	10	5	10	9
2	13	5	13	11	13	10	12	11
3	8	8	5	6	6	8	6	8
4	6	5	7	5	8	5	7	4
5	8	7	11	5	6	6	12	10

in VOT and syntax comprehension, such as temperature, fatigue, and alcohol intoxication, can be easily ruled out. Although temperature occasionally fell to −40°C, the test sessions were conducted at Base Camp at moderate temperatures (15–25°C) and at Camps Two and Three in tents by climbers wearing down climbing suits or immediately outside their tents in sunny, warm, windless conditions. General fatigue cannot account for the observed pattern because the tests administered upon return to Base Camp were conducted just after the climbers had completed the most strenuous part of the climb (a minimum of 16 hours in continuous activity ascending from Camp Four to the summit and back before descending to Camp Two and to Base Camp), yet VOT separation width was larger and RITLS response time was shorter than at the higher Camps. Alcohol intoxication affects speech production to such an extent that speech measurements have been proposed to evaluate sensory and motor impairments due to alcohol consumption (Brenner & Cash, 1991; Pisoni & Martin, 1989). However, no alcohol was ingested by any of the subjects while they were at Camps Two and Three or before the test sessions at Base Camp.

Hypoxia, selectively affecting some brain structures, remains the most likely cause of our findings. Acclimatization would have been minimal at testing time because subjects were tested at Camps Two and Three upon arrival at those altitudes. Furthermore, acclimatization is incomplete at these altitudes; it is unclear whether any improvement occurs for longer stays above 6,000 m (West, 1985).

Exposure to extreme altitude does not appear to affect all aspects of cognitive behavior to the same degree. Long-term memory, for example, was not affected in a previous study at Everest Camp Two when it was not combined with learning (Nelson et al., 1990).[5] The pattern of deficits noted at extreme altitude is, therefore, consistent with other studies that indicate that the neural bases of long term memory and the lexicon appear to be dissociable from those regulating speech motor control and syntax. Speech motor control and syntax, for example, are preserved in Alzheimer's disease which affects lexical ability and memory (Kempler, Curtiss, & Jackson, 1987; Kempler, 1988). In contrast, long term memory and lexical ability are preserved in non-demented Parkinson's patients who show syntax comprehension and VOT motor control deficits (Lieberman et al., 1992).

It is possible that the VOT and syntax comprehension deficits we report here have a similar neurological basis to Parkinson's disease, i.e., they may reflect the degradation of basal ganglia pathways to prefrontal cortex. Similar patterns of deficits have been found in patients with Parkinson's disease, where the main pathological findings are compromised cells in the basal ganglia, particularly in the substantia nigra, and throughout the dopaminergic pathways in the lentiform and caudate nuclei to the prefrontal cortex (Cummings, 1993; Parent, 1986). Non-demented Parkinson's patients tested with these tasks have shown small decrements in digit span and confrontation naming performance in contrast to high RITLS error rates. They also show speech motor control deficits evidenced in VOT measurements (Lieberman et al., 1992). Although the extent of the deficits of Parkinson's patients is much larger than that of our subjects the pattern is strikingly similar. The severity of the deficits cannot be expected to be comparable, because Parkinson's patients may have profound impairments whereas individuals fit enough to climb Mt. Everest cannot possibly be very severely impaired.

An important theoretical implication of the correlation between the impairments in speech motor control and in sentence comprehension is that it is not consistent with the "syntax module" advocated by some researchers (e.g., Pinker, 1994; Wilkins & Wakefield, 1994). That is, unless speech motor control and syntactic deficits are dissociable, the argument for a brain structure specialized for syntactic processing is empirically unsupported. The observed correlation may reflect the preadaptive role of speech in the evolution of syntax (Lieberman, 1991, 1992).

Apart from its theoretical importance for neural and linguistic theories, this correlation can also have practical applications. Cognitive impairments have been frequently observed in high altitude climbing and flying and have often led to accidents due to improper evaluation of danger or other poor judgement (Ward et al., 1989; Nelson et al., 1990). If

[5]It must be noted that other studies have found memory deficits (e.g., Regard et al., 1989, 1991; Cavaletti et al., 1987), but not with identical tasks; the length of exposure to high altitude and the actual altitude also have varied between studies. Furthermore, as Nelson et al. (1990) have pointed out, memory testing in which impairments were observed had included a learning component.

syntax comprehension deficits, such as those we observed, are a good index of the other cognitive impairments then our findings suggest that speech motor control, as measured through VOT separation width, can provide an estimate of the extent of the impairment. Thus, we can construct a remote monitoring system to automatically measure VOT separation width in naturally occurring speech (e.g., for communication) to assess neural functioning of personnel involved in hazardous situations where the consequences of error can be grave, such as aeronautics, spaceflight, and flight control.

Initial applications will certainly include situations not only of hypoxic hypoxia but also of anaemic hypoxia (e.g., from carbon monoxide intoxication), because the same brain structures are again the most sensitive (globus pallidus in the basal ganglia, hippocampus, and parts of the substantia nigra; Brierley, 1976; Laplane, Levasseur, Pillon, Dubois, Baulac, Mazoyer, Dinh, Sette, Danze, & Baron, 1989) and perhaps of alcohol intoxication, if a similar relationship holds between speech motor control and the extent of the impairment. Furthermore, it will be possible to use similar methods to remotely evaluate the treatment of neurodegenerative diseases such as Parkinson's. A patient would just make a phonecall and an automated speech analysis system could aid the physician to evaluate their progress and adjust the treatment accordingly.

Conclusion

We have found correlated deficits in speech motor control and syntax comprehension in five subjects climbing Mt. Everest. These deficits were more pronounced at higher altitudes; no deficits were found in other cognitive tasks, yielding a pattern similar to that found in Parkinson's patients. We argue that the impairments are due to disruption of basal ganglia pathways to prefrontal cortex caused by hypoxia, in agreement with neuropathological findings on the vulnerability of brain structures. The theoretical implications of our findings are in favor of a basal ganglia involvement in many functions besides motor control and not consistent with a specialized syntax module in the brain.

The practical applications of our study are also very important. Previous studies have identified various cognitive impairments in climbers at high altitudes, but have not offered a practical way to remotely assess their neural functioning. While symptoms of acute mountain sickness (Regard et al., 1991) and ventilatory response (Ward et al., 1989) have also been reported to correlate with cognitive performance, the former is not useful in prevention and the latter is far less easy to monitor than VOT separation width in speech that is readily available from the communications channels.

More research is necessary to refine and validate the procedure we propose, and to build a compact automatic speech analysis system for remote monitoring. Such systems may become indispensable in various situations where monitoring crew behavior is critical, as well as in the day-to-day treatment of Parkinson's disease.

References

Baum, S. (1988). Syntactic processing in agrammatism: Evidence from lexical decision and grammaticality judgment tasks. *Aphasiology, 2*, 117–135.

Baum, S., Blumstein, S. E., Naeser, M. A., & Palumbo, C. L. (1990). Temporal dimensions of consonant and vowel production: And acoustic and CT scan analysis of aphasic speech. *Brain and Language, 39*, 33–56.

Brenner, M. & Cash, J. R. (1991). Speech analysis as an index of alcohol intoxication. *Aviation, Space, and Environmental Medicine, 62*, 893–898.

Brierley, J. B. (1976). Cerebral hypoxia. In W. Blackwood & J. A. N. Corsellis (Eds.), *Greenfield's Neuropathology*, chap. 2, pp. 43–85. Chicago: Year Book Medical Publishers.

Cavaletti, G., Moroni, R., Garavaglia, P., & Tredici, G. (1987). Brain damage after high-altitude climbs without oxygen. *The Lancet, 1*, 101.

Clark, C. F., Heaton, R. K., & Wiens, A. N. (1983). Neurophysiological functioning after prolonged high altitude exposure in mountaineering. *Aviation, Space, and Environmental Medicine, 54*(3), 202–207.

Cummings, J. L. (1993). Frontal-subcortical circuits and human behavior. *Archives of Neurology, 50*, 873–880.

Deecke, L., Goode, R. C., Whitehead, G., Johnson, W. H., & Bryce, D. P. (1973). Hearing under respiratory stress: Latency changes of the human auditory evoked response during hyperventilation, hypoxia, asphyxia, and hypercapnia. *Aerospace Medicine, 44*(10), 1106–1111.

Engen, E. & Engen, T. (1983). *Rhode Island Test of Language Structure.* Baltimore: University Park Press.

Fowler, B. & Kelso, B. (1992). The effects of hypoxia on components of the human-related potential and relationship to reaction time. *Aviation, Space, and Environmental Medicine, 63*, 510–516.

Grossman, M., Carvell, S., Stern, M. B., Vernon, G., & Hurtig, H. I. (1991). Sentence comprehension and praxis deficits in Parkinson's disease. *Neurology, 41*, 1620–1626.

Hornbein, T. F., Townes, B. D., Schoene, R. B., Sutton, J. R., & Houston, C. S. (1989). The cost to the central nervous system of climbing to extremely high altitude. *The New England Journal of Medicine, 321*(25), 1714–1719.

Jason, G. W., Pajurkova, E. M., & Lee, R. G. (1989). High-altitude mountaineering and brain function: Neurophysiological testing of members of a Mount Everest expedition. *Aviation, Space, and Environmental Medicine, 60*, 170–173.

Kempler, D. (1988). Lexical and pantomime abilities in Alzheimer's disease. *Aphasiology, 2*, 147–159.

Kempler, D., Curtiss, S., & Jackson, C. (1987). Syntactic presevation in Alzheimer's disease. *Journal of Speech and Hearing Research, 30*, 343–350.

Kennedy, R. S., Dunlap, W. P., Banderet, L. E., Smith, M. G., & Houston, C. S. (1989). Cognitive performance deficits in a simulated climb of Mount Everest: Operation Everest II. *Aviation, Space, and Environmental Medicine*, *60*, 99–104.

Lange, K. W., Robbins, T. W., Marsden, C. D., James, M., Owen, A. M., & Paul, G. M. (1992). L-Dopa withdrawal in Parkinson's disease selectively impairs cognitive performance in tests sensitive to frontal lobe dysfunction. *Psychopharmacology*, *107*, 394–404.

Laplane, D., Levasseur, M., Pillon, B., Dubois, B., Baulac, M., Mazoyer, B., Dinh, S. T., Sette, G., Danze, F., & Baron, J. C. (1989). Obsessive-compulsive and other behavioral changes with bilateral basal ganglia lesions. *Brain*, *112*, 699–725.

Lieberman, P. (1991). *Uniquely human: The evolution of speech, thought, and selfless behavior*. Cambridge, Mass: Harvard University Press.

Lieberman, P. (1992). Could an autonomous syntax module have evolved?. *Brain and Language*, *43*, 768–774.

Lieberman, P., Kako, Ė., Friedman, J., Tajchman, G., Feldman, L. S., & Jiminez, E. B. (1992). Speech production, syntax comprehension, and cognitive deficits in Parkinson's disease. *Brain and Language*, *43*, 169–189.

Lisker, L. & Abramson, A. S. (1964). A cross language study of voicing in initial stops: Acoustical measurements. *Word*, *20*, 384–442.

Metter, E. J., Kempler, D., Jackson, C. A., Hanson, W. R., Mazziota, J. C., & Phelps, M. E. (1989). Cerebral glucose metabolism in Wernicke's, Broca's, and conduction aphasia. *Archives of Neurology*, *46*, 27–34.

Metter, E. J., Riege, W. H., Hanson, W. R., Phelps, M. E., & Kuhl, D. E. (1984). Local cerebral metabolic rates of glucose in movement and language disorders from positron tomography. *American Journal of Physiology*, *246*, R897–900.

Nelson, T. O., Dunlosky, J., White, D. M., Steinberg, J., Townes, B. D., & Anderson, D. (1990). Cognition and metacognition at extreme altitudes on Mount Everest. *Journal of Experimental Psychology: General*, *119*(4), 367–374.

Oelz, O., Regard, M., Wichmann, W., Valavanis, A., Witztum, A., Brugger, P., Cerretelli, P., & Landis, T. (1990). Cognitive impairment, neurologic performance and MRI after repeat exposure to extreme altitude. In J. R. Sutton, G. Coates, & J. E. Remmers (Eds.), *Hypoxia: The Adaptations*, chap. 33, pp. 206–209. Toronto: Decker.

Parent, A. (1986). *Comparative neurobiology of the basal ganglia*. New York: John Wiley & Sons.

Parkinson's study group (1989). The DATATOP series. *Archives of Neurology*, *46*, 1052–1060.

Pinker, S. (1994). *The Language Instinct*. New York: William Morrow.

Pisoni, D. B. & Martin, C. S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. *Alcoholism: Clinical and Experimental Research*, *13*(4), 577–587.

Regard, M., Landis, T., Casey, J., Maggiorini, M., Bärtsch, P., & Oelz, O. (1991). Cognitive changes at high altitude in healthy climbers and in climbers developing acute mountain sickness. *Aviation, Space, and Environmental Medicine*, *62*, 291–295.

Regard, M., Oelz, O., Brugger, P., & Landis, T. (1989). Persistent cognitive impairment in climbers after repeated exposure to extreme altitude. *Neurology*, *39*, 210–213.

Townes, B. D., Hornbein, T. F., Schoene, R. B., Sarnquist, F. H., & Grant, I. (1984). Human cerebral function at extreme altitude. In J. B. West (Ed.), *High altitude and man*, chap. 3, pp. 31–36. Bethesda, Maryland: American Psychological Society.

Ward, M. P., Milledge, J. S., & West, J. B. (1989). *High altitude medicine and physiology*, chap. 17, pp. 311–325. Philadelphia: University of Pennsylvania Press.

West, J. B. (1985). *Everest: The testing place*. New York: McGraw-Hill.

Wilkins, W. K. & Wakefield, J. (1994). Brain evolution and neurolinguistic preconditions. *Behavioral and Brain Sciences*.

SPEECH PERCEPTION AS A TALKER-CONTINGENT PROCESS

Lynne C. Nygaard, Mitchell S. Sommers, and David B. Pisoni
Indiana University

Abstract—*To determine how familiarity with a talker's voice affects perception of spoken words, we trained two groups of subjects to recognize a set of voices over a 9-day period. One group then identified novel words produced by the same set of talkers at four signal-to-noise ratios. Control subjects identified the same words produced by a different set of talkers. The results showed that the ability to identify a talker's voice improved intelligibility of novel words produced by that talker. The results suggest that speech perception may involve talker-contingent processes whereby perceptual learning of aspects of the vocal source facilitates the subsequent phonetic analysis of the acoustic signal.*

During the perception of speech, listeners must extract stable phonetic percepts from acoustic signals that are highly variable. Variations in talker characteristics, in particular, have been shown to produce profound effects on the acoustic realization of speech sounds (Nearey, 1978; Peterson & Barney, 1952). Traditionally, models of speech perception have characterized variation in the acoustic speech signal as a perceptual problem that perceivers must solve (Shankweiler, Strange, & Verbrugge, 1976). Listeners are thought to contend with variation in speech signals due to talker differences through a compensatory process in which speech sounds are normalized with reference to specific voice characteristics. According to a strict interpretation of this view, information about a talker is stripped away during the perception of speech to arrive at the abstract, canonical linguistic units that are presumed to be the basic building blocks of perception (Halle, 1985; Joos, 1948; Summerfield & Haggard, 1973).

Address correspondence to Lynne C. Nygaard, Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405; e-mail: lnygaard@indiana.edu.

Unfortunately, this standard view of talker normalization begs the question of how the processing of a talker's voice is related to the perception of the phonetic content of speech. Although a talker's voice carries important information about the social and physical aspects of that talker into the communicative setting (Laver & Trudgill, 1979), the encoding of voice-specific information for the identification and discrimination of talkers has generally been considered to be a problem quite separate from apprehending the linguistic content of an utterance. On the one hand, researchers have investigated the ability of listeners to explicitly recognize and discriminate familiar and unfamiliar voices (e.g., Legge, Grossmann, & Pieper, 1984; Van Lancker, Kreiman, & Emmorey, 1985). In this case, the speech signal is viewed simply as a carrier of talker information. On the other hand, research in speech perception has been devoted to studying the linguistic content of speech—either entirely independently of any variability in talker or source characteristics or from the point of view that variation due to changes in talkers is noise that must be normalized or discarded quickly in order to recover the linguistic content of an utterance. Consequently, the emphasis in speech perception research has been on identifying and defining short-term, presumably automatic adaptations to differences in source characteristics (Garvin & Ladefoged, 1963; Johnson, 1990; Ladefoged & Broadbent, 1957; Miller, 1989; Nearey, 1989).

The theoretical and empirical dissociation of the encoding of talker characteristics and the processing of the phonetic content of an utterance assumes that the analysis of these two kinds of information is independent (Laver & Trudgill, 1979). Only recently has this assumption been questioned on the basis of a growing body of research demonstrating effects of talker variability on both perceptual (Mullennix & Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989; Summerfield & Haggard, 1973) and memory (Goldinger, Pisoni, & Logan, 1991; Martin, Mullennix, Pisoni, & Summers, 1989) processes. For example, using a continuous recognition memory procedure, Palmeri, Goldinger, and Pisoni (1993) recently found that specific voice information was retained in memory along with item information, and these attributes were found to aid later recognition memory. These findings suggest that talker information may not be discarded in the process of speech perception, but rather variation in a talker's voice may become part of a rich and highly detailed representation of the speaker's utterance.

Although previous experiments have demonstrated that short-term adjustments may occur in the analysis of speech produced by different talkers (Ladefoged & Broadbent, 1957) and that talker information may be retained in long-term memory, the question remains whether the talker information that is retained in memory has any relationship to the ongoing analysis of linguistic content during the perception of speech. The purpose of the present experiment was to address this question by determining if differences in a listener's familiarity with a vocal source have any effect on the encoding of the phonetic content of a talker's utterance. To accomplish this, we asked two groups of listeners explicitly to learn to recognize the voices of 10 talkers over a 9-day period. At the end of the training period, we evaluated the role of talker recognition on the perception of spoken words to determine if the ability to identify a talker's voice was independent of phonetic analyses. It should be noted that independence between talker recognition and phonetic analysis is implicitly assumed by all current theoretical accounts of speech perception (Fowler, 1986; Liberman & Mattingly, 1985; McClelland & Elman, 1986; Stevens & Blumstein, 1978). If learning

Psychological Science, Vol. 5, No. 1, Nygaard, L.C., Sommers, M.S., & Pisoni, D.B., Speech perception as a talker-contingent process, pp. 42-46, (1994). Reprinted with the permission of Cambridge University Press.

to identify a talker's voice is found to affect subsequent word recognition performance, the mechanisms responsible for the encoding of talker information would seem to be linked directly to those that underlie phonetic perception. Establishing such a link would require a fundamental change in present conceptualizations of the nature of mechanisms contributing to speech perception.

METHOD

Subjects

Subjects were 38 undergraduate and graduate students at Indiana University. Nineteen subjects served in each condition—experimental and control. All subjects were native speakers of American English and reported no history of a speech or hearing disorder at the time of testing. The subjects were paid for their services.

Stimulus Materials

Three sets of stimuli were used in this experiment. All were selected from a data base of 360 monosyllabic words produced by 10 male and 10 female talkers. Word identification tests in quiet showed greater than 90% intelligibility for all words. In addition, all words were rated to be highly familiar (Nusbaum, Pisoni, & Davis, 1984). The stimuli were originally recorded on audiotape and digitized at a sampling rate of 10 kHz on a PDP 11/34 computer using a 12-bit analog-to-digital converter. The root mean squared (RMS) amplitude levels for all words were digitally equated.

Procedure

Training

Two groups of 19 listeners each completed 9 days of training to familiarize themselves with the voices of 10 talkers. Listeners were asked to learn to recognize each talker's voice and to associate that voice with one of 10 common names (see Lightfoot, 1989). Digitized stimuli were presented using a 12-bit digital-to-analog converter and were low-pass filtered at 4.8 kHz. Stimuli were presented to listeners over matched and calibrated TDH-39 headphones at approximately 80 dB SPL (sound pressure level).

On each of the 9 training days, both groups of listeners completed three different phases. The first phase consisted of a familiarization task. Five words from each of the 10 talkers were presented in succession to the listeners. Subjects then heard a 10-word list composed of 1 word from each talker in succession. Each time a token was presented to the listeners, the name of the appropriate talker was displayed on a computer screen. Listeners were asked to listen carefully to the words presented and to attend specifically to the talker's voice so they could learn the name.

The second phase of training consisted of a recognition task in which subjects were asked to identify the talker who had produced each token. The 100 words used did not overlap with those used in the first phase. Ten words from each of the 10 talkers were presented in random order to listeners who were asked to recognize each voice by pressing the appropriate button on a keyboard. The keys were labeled with 10 names. Keys 1 through 5 were labeled with male names; Keys 6 through 10 were labeled with female names. On each trial, after all subjects had entered their responses, the correct name appeared on the computer screen.

After subjects completed two repetitions of the first two phases of training, we administered a test phase on each day. As in the second training phase, 10 words from each of the 10 talkers were presented in random order. Subjects were asked to indicate who each speaker was by pressing on a keyboard the button corresponding to the appropriate name. However, feedback was not given.

Although the words used in the test phase were drawn from the same 100 words used in the second training phase, on each day of training subjects never heard the same item produced by the same talker in both the test and the training phase. In addition, training stimuli were reselected from the data base on each day so that subjects never heard the same word produced by the same talker in training. This training procedure was designed to expose listeners to a diverse set of tokens from each of the talkers.

Generalization

On the 10th day of the experiment, both groups of subjects completed a generalization test. One hundred new words produced by each of the 10 familiar talkers were used. As in the test phase used during training, 10 words from each of the 10 talkers were presented in random order. Subjects were asked to name the talker on each trial. No feedback was given. Thus, the generalization test was identical to the training test phase except that listeners had never heard any of the words before.

Word intelligibility

In addition to the generalization test, we administered a speech intelligibility test in which subjects were asked to identify words presented in noise. In this transfer task, 100 novel words were presented at either 80, 75, 70, or 65 dB (SPL) in continuous white noise low-pass filtered at 4.8 kHz and presented at 70 dB (SPL), yielding four signal-to-noise ratios: +10, +5, 0, and −5. Equal numbers of words were presented at each of the four signal-to-noise ratios. In this test, subjects were simply asked to identify the word itself (rather than explicitly recognize the talker's voice) by typing the word on a keyboard. Subjects in the experimental condition were presented with words produced by the 10 talkers they had learned in the training phase. Subjects in the control condition were presented with words produced by 10 new talkers they had not heard in the training phases.

RESULTS AND DISCUSSION

Training

Most subjects showed continuous improvement across the 9 days in their ability to recognize talkers from isolated words. However, individual differences were found in performance. Consequently, we selected a criterion of 70% correct for talker recognition on the last day of training for inclusion in the experiment. Our rationale for choosing this criterion was simply that to determine whether learning a talker's voice affects perceptual processing, we needed to ensure we had identified a group of subjects who did, in fact, learn to recognize the talkers' voices from isolated words. On the basis of this criterion, 9 subjects from each training group were included in the final analysis.[1] Both groups of listeners identified talkers consistently above chance even on the 1st day of training, and performance rose to nearly 80% correct by the last day of training. A repeated measures analysis of variance (ANOVA) with learning and days of training as factors showed a significant main effect of day of training, $F(9, 144) = 73.55$, $p < .0001$, but no difference between the two groups over days of training, $F(1, 16) = 0.14$, $p > .7$.

Generalization

The generalization test showed almost identical recognition of voices from novel words on the 10th day as on the final day of training. The implication of this result is that listeners acquired detailed knowledge about the talkers' voices that was not necessarily dependent on the specific words that carried that information. In other words, the perceptual learning that took place in the course of the nine training sessions was not dependent on the training stimuli but rather readily generalized to novel utterances produced by the same set of talkers.

Word Intelligibility

Figure 1 shows the percentage of correct word identification as a function of signal-to-noise ratio for both groups of trained subjects. As expected, identification performance decreased from the $+10$ to the -5 signal-to-noise ratio for both groups. However, subjects tested with words produced by familiar voices were significantly better in recognizing novel words at each signal-to-noise ratio than were subjects tested with unfamiliar

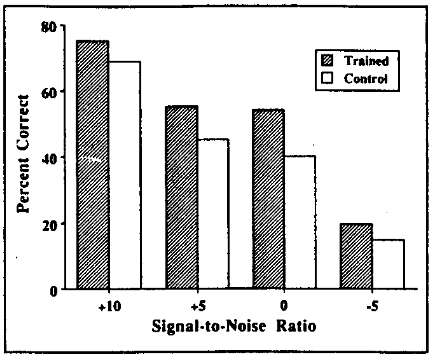


Fig. 1. Mean intelligibility of words presented in noise for trained and control subjects. Trained, or experimental, subjects were trained with one set of talkers and tested with words produced by these familiar talkers. Control subjects were trained with one set of talkers and tested with words produced by a novel set of talkers. Percentage of correct word recognition is plotted at each signal-to-noise ratio.

voices.[2] A repeated measures ANOVA with training and signal-to-noise ratio as factors revealed highly significant main effects of both signal-to-noise ratio, $F(3, 48) = 173.27$, $p < .0001$, and experimental condition (experimental vs. control group), $F(1, 16) = 13.62$, $p < .002$.[3]

To ensure that the overall intelligibility of the two sets of voices did not differ, two additional groups of 18 untrained subjects who were not familiar with either set of talkers were given the same word intelligibility test. One untrained control group received the stimulus tokens produced by the talkers who were used in the training phase; the other untrained control group received the stimulus tokens from the talkers who were presented to the trained control group in the intelligibility test. Identification performance for the trained and untrained control groups did not differ. A separate repeated measures ANOVA including the two untrained and the one trained control conditions revealed a significant main effect of signal-to-noise ratio, $F(3, 102) = 221.38$, $p < .001$, but no significant main effect of control condition, $F(2, 34) = 0.16$, $p > .9$. This finding confirms that the difference in performance between the experimental group and the trained control subjects was not due to inherent differences in the intelligibility of the voices or the words used.

GENERAL DISCUSSION

The present study found that voice recognition and processing of the phonetic content of a linguistic utterance were not independent. Listeners who learned to recognize a set of talkers apparently encoded and retained in long-term memory talker-specific information that facilitated the subsequent perceptual analysis and identification of novel words produced by the same talkers. These findings provide the first demonstration that experience identifying a talker's voice facilitates perceptual processing of the phonetic content of that speaker's novel utterances. Not only does the perceptual learning that results from the talker recognition task generalize to the recognition of familiar voices producing novel words, but that learning also transfers to a completely different task involving the perceptual analysis of the phonetic content of novel words produced by the same talkers in a speech intelligibility test. Listeners who were presented with one set of voices but were tested with another set of voices failed to show any benefit from the experience gained by learning to recognize those voices explicitly. Only experience with the specific voices used in the intelligibility test facilitated the phonetic processing of novel words. The implication of this result is that phonetic perception and spoken word recognition appear to be affected by knowledge of specific information about a talker's voice. Experience with specific acoustic attributes of a talker's voice appears to facilitate the analysis of spoken words.

1. It should be noted that the task of learning to identify voices from isolated words is extremely difficult (see Williams, 1964). Therefore, it was necessary to set a somewhat arbitrary training period and then select subjects who had learned to our criterion by the end of that period. Given additional training with isolated words or with sentences or larger passages of speech, a greater percentage of our subjects would have reached a criterion level of performance.

2. It should be noted that the subjects who did not meet the criterion of 70% correct voice identification on Day 9 of training were also tested in the word recognition task. Among these "poor" learners, subjects who received words produced by talkers previously heard in training showed no advantage over subjects who received words produced by talkers not heard previously. This finding suggests that simple exposure to the voices heard in training was not sufficient for listeners to obtain the perceptual learning necessary for improved word recognition ability.

3. Four items from the control condition were eliminated from the overall analyses. After the experiment had been run, these items were found to be mispronounced.

Our results are consistent with the view that the learning that occurs when listeners are trained to recognize and identify talkers' voices involves the modification of the procedures or perceptual operations necessary for the extraction of voice information from the speech signal (Kolers, 1976; Kolers & Roediger, 1984). That is, over time during training, listeners may learn to attend to and modify the specific perceptual operations used to analyze and encode each talker's voice during perception, and it is these talker-specific changes that are retained in memory. This procedural knowledge would then allow listeners to more efficiently analyze novel words produced by familiar talkers. We believe this situation may be very similar to the case of reading and remembering inverted text. Kolers and Ostry (1974) found that the operations necessary to read inverted text were retained in long-term memory and facilitated subsequent tasks involving reading inverted text. The type of detailed procedural knowledge that Kolers and Ostry described may be responsible for subjects' superior performance in identifying words spoken by familiar talkers in the present experiment.

The present findings demonstrate that the process that contends with variation in talker characteristics can be modified by experience and training with a specific talker's voice. The interaction of learning to identify a talker's voice and processing the phonetic content of a talker's utterance suggests that the speech perception mechanism is susceptible to general processes of perceptual learning and attention. Thus, the processing of a talker's voice may demand time and resources if the voice is unfamiliar to a listener (Martin et al., 1989; Mullennix et al., 1989; Summerfield & Haggard, 1973), but may become much more efficient if the voice can be identified as familiar (Lightfoot, 1989). The fact that the speech-processing system is susceptible to such modification argues against a strictly modular view of phonetic processing (Fodor, 1983; Liberman & Mattingly, 1985). We find that encoding of a talker's voice interacts extensively with the analysis of spoken words. If word recognition were a separable process or module distinct from voice recognition, then training listeners to identify voices should have no effect on speech intelligibility. However, this experiment shows that learning to identify voices does facilitate perceptual analysis of words produced by those voices, indicating that encoding of voice characteristics and the perception of speech are highly integral processes that work together to perceptually organize the interleaved talker and linguistic information present in the acoustic signal.

Finally, this study provides the first direct demonstration of the role of long-term memory and perceptual learning of source characteristics in speech perception and spoken word recognition. The perceptual learning acquired through a task involving explicit identification and labeling of talkers' voices was found to transfer to an entirely different task involving the perception of the linguistic content of an utterance. It appears that the analysis involved when different voices are encountered in the perceptual process is not limited to short-term, on-line normalization, as supposed by most current theories of speech perception, but rather is a highly modifiable process that is subject to the perceptual learning of talker-specific information. Indeed, the present findings suggest that the phonetic coding of speech is carried out in a talker-contingent manner. Phonetic perception and spoken word recognition appear to be integrally related to knowledge of characteristics of a talker's vocal tract and, consequently, attributes of a talker's voice.

Acknowledgments—This research was supported by Training Grant DC-00012-13 and Research Grant DC-0111-16 from the National Institutes of Health to Indiana University. We thank Carol A. Fowler, Robert E. Remez, Peter D. Eimas, Scott E. Lively, and Stephen D. Goldinger for their helpful comments and suggestions on an earlier version of this manuscript.

REFERENCES

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*, 3–28.

Garvin, P.L., & Ladefoged, P.L. (1963). Speaker identification and message identification in speech recognition. *Phonetica, 9*, 193–199.

Goldinger, S.D., Pisoni, D.B., & Logan, J.S. (1991). On the nature of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 152–162.

Halle, M. (1985). Speculations about the representation of words in memory. In V.A. Fromkin (Ed.), *Phonetic linguistics* (pp. 101–104). New York: Academic Press.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America, 88*, 642–654.

Joos, M.A. (1948). Acoustic phonetics. *Language, 24*(Suppl. 2), 1–136.

Kolers, P.A. (1976). Pattern analyzing memory. *Science, 191*, 1280–1281.

Kolers, P.A., & Ostry, D.J. (1974). Time course of loss of information regarding pattern analyzing operations. *Journal of Verbal Learning and Verbal Behavior, 13*, 599–612.

Kolers, P.A., & Roediger, H.L., III. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior, 23*, 425–449.

Ladefoged, P., & Broadbent, D.E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*, 98–104.

Laver, J., & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K.R. Scherer & H. Giles (Eds.), *Social markers in speech* (pp. 1–32). Cambridge, England: Cambridge University Press.

Legge, G.E., Grossmann, C., & Pieper, C.M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 298–303.

Liberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36.

Lightfoot, N. (1989). *Effects of talker familiarity on serial recall of spoken word lists* (Research on Speech Perception Progress Report No. 15). Bloomington: Indiana University.

Martin, C.S., Mullennix, J.W., Pisoni, D.B., & Summers, W.V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 676–681.

McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86.

Miller, J.D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America, 85*, 2114–2134.

Mullennix, J.W., & Pisoni, D.B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*, 379–390.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85*, 365–378.

Nearey, T. (1978). *Phonetic features for vowels*. Bloomington: Indiana University Linguistics Club.

Nearey, T. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85*, 2088–2113.

Nusbaum, H.C., Pisoni, D.B., & Davis, D.K. (1984). *Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words* (Research on Speech Perception Progress Report No. 10). Bloomington: Indiana University.

Palmeri, T.J., Goldinger, S.D., & Pisoni, D.B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 309–328.

Peterson, G.E., & Barney, H.L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175–184.

Shankweiler, D.P., Strange, W., & Verbrugge, R.R. (1976). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, knowing: Toward an ecological psychology* (pp. 315–345). Hillsdale, NJ: Erlbaum.

Stevens, K.N., & Blumstein, S.E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64,* 1358–1368.

Summerfield, Q., & Haggard, M.P. (1973). *Vocal tract normalisation as demonstrated by reaction times* (Report on Research in Progress in Speech Perception No. 2). Belfast, Northern Ireland: Queen's University of Belfast.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I. Recognition of backward voices. *Journal of Phonetics, 13,* 19–38.

Williams, C.E. (1964). The effects of selected factors on the aural identification of speakers (Section III of Report EDS-TDR-65-153). Hanscom Field, MA: Electronic Systems Division, Air Force Systems Command.

(Received 12/15/92; Revision accepted 6/7/93)

REPRINTED FROM:

SPEECH
COMMUNICATION

Speech Communication 13 (1993) 109–125
North-Holland

Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning *

David B. Pisoni

Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405, USA

Received 25 January 1993
Revised 19 May 1993

N·H

NORTH-HOLLAND
AMSTERDAM – LONDON – NEW YORK – TOKYO

Speech Communication 13 (1993) 109–125
North-Holland

Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning *

David B. Pisoni

Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405, USA

Received 25 January 1993
Revised 19 May 1993

Abstract. This paper summarizes results from recent studies on the role of long-term memory in speech perception and spoken word recognition. Experiments on talker variability, speaking rate and perceptual learning provide strong evidence for implicit memory for very fine perceptual details of speech. Listeners apparently encode specific attributes of the talker's voice and speaking rate into long-term memory. Acoustic–phonetic variability does not appear to be "lost" as a result of phonetic analysis. The process of perceptual normalization in speech perception may therefore entail encoding of specific instances or "episodes" of the stimulus input and the operations used in perceptual analysis. These perceptual operations may reside in a "procedural memory" for a specific talker's voice. Taken together, the present set of findings are consistent with non-analytic accounts of perception, memory and cognition which emphasize the contribution of episodic or exemplar-based encoding in long-term memory. The results from these studies also raise questions about the traditional dissociation in phonetics between the linguistic and indexical properties of speech. Listeners apparently retain non-linguistic information in long-term memory about the speaker's gender, dialect, speaking rate and emotional state, attributes of speech signals that are not traditionally considered part of phonetic or lexical representations of words. These properties influence the initial perceptual encoding and retention of spoken words and therefore should play an important role in theoretical accounts of how the nervous system maps speech signals onto linguistic representations in the mental lexicon.

Zusammenfassung. Dieser Artikel faßt die Ergebnisse von kürzlich durchgeführten Untersuchungen über die Aufgabe des langfristigen Gedächtnisses bei der Erkennung von Sprache und Noten. Diese Erfahrungen über die Veränderlichkeit der Redner, des Sprachflußes und des aufnehmenden Lernens zeigen deutlich das Bestehen eines Gedächtnisses für die sehr feinen, aufnehenden Details der Sprache. Anscheinend codieren die Hörer die spezifischen Attribute der Stimme des Redners und seines Sprachflusses in einem langfristigen Gedächtnis. Die akustisch–phonetische Veränderlichkeit scheint nach der phonetischen Analyse nicht verloren zu sein. Der bei der Spracherkennung eingesetzte Prozessor für die Normalisierung der Aufnahme scheint somit eine Codierung der spezifischen Ereignisse oder "Episoden" des Eingangsstimulus und der aufnehmenden Analyse zu enthalten. Diese Vorgänge der Aufnahme könnten innerhalb eines "Verfahrenspeichers" für jede gegebene Rednerstimme durchgeführt werden. Global gesehen stimmen die aktuellen Beobachtungen mit den nicht analytischen Beschreibungen der Aufnahme, des Gedächtnisses und der Erkennung überein, was die Bedeutung einer episodischen Codierung oder eine Codierung "anhand von Beispielen" für das langfristige Gedächtnis zeigt. Die Ergebnisse dieser Untersuchungen unterstreichen auch die Frage der traditionellen Trennung in der phonetik zwischen den sprachlichen Eigenschaften und der Indexierung der Sprache. Anscheinend behalten die Hörer im langfristigen Gedächtnis nicht sprachliche Informationen über die Art des Redners, seinen Dialekt, seinen Sprachfluß und seinen emotionellen Zustand; alles Attribute des Sprachsignals, die allgemein nicht als relevant für die phonetischen oder lexikalen Darstellungen der Worte berücksichtigt werden. Diese Eigenschaften beeinflußen die ursprüngliche Codierung der Aufnahme und die Erinnerung der gesprochenen Worte und spielen somit eine bedeutende Rolle bei den theoretischen Hypothesen über die Art, wie das Nervensystem die Sprachsignale mit den sprachlichen Darstellungen im Gedächtnis in Verbindung bringt.

* Those of us who work in the field of human speech perception owe a substantial intellectual debt to Professor Hiroya Fujisaki, who has contributed in many important ways to our current understanding of the speech mode and the underlying perceptual mechanisms. His theoretical and empirical work in the late 1960's brought the study of speech perception directly into the main stream of cognitive psychology (Fujisaki and Kawashima, 1969). In particular, his pioneering research and modeling efforts on categorical perception inspired a large number of empirical studies on issues related to coding processes and the contribution of short-term memory to speech perception and categorization. This paper is dedicated to Professor Fujisaki, one of the great pioneers in the field of speech research and spoken language processing.

Elsevier Science Publishers B.V.

Résumé. Cet article résume les résultats d'études récentes sur le rôle de la mémoire à long-terme en perception de parole et en reconnaissance de mots. Ces expériences sur la variabilité des locuteurs, le débit de parole et l'apprentissage perceptif fournissent des preuves fortes en faveur de l'existence d'une mémoire implicite pour les détails perceptifs très fins de la parole. Apparemment, les auditeurs encodent des attributs spécifiques de la voix du locuteur et de son débit dans une mémoire à long-terme. La variabilité acoustico–phonétique ne semble pas être "perdue" à l'issue de l'analyse phonétique. Le processus de normalisation perceptive en oeuvre lors de la perception de parole semble donc comporter un encodage d'événements ou "épisodes" spécifiques du stimulus d'entrée et les opérations d'analyse perceptive. Ces opérations perceptives pourraient être effectuées au sein d'une "mémoire procédurale" pour chaque voix de locuteur donnée. Globalement, l'ensemble actuel d'observations est en accord avec les présentations non-analytiques de la perception, de la mémoire et de la cognition qui mettent l'accent sur la contribution d'un encodage épisodique ou "à base d'exemples" au processus de mémoire à long terme. Les résultats de ces études posent également la question des dissociations tradionnelles en phonétique entre les propriétés linguistiques et d'indexation de la parole. Les auditeurs retiennent apparemment, dans la mémoire à long-terme, des informations non-linguistiques sur le genre du locuteur, son dialecte, son débit de parole et son état émotionnel. attributs du signal de parole qui ne sont généralement pas considérés comme relevant des représentations phonétiques ou lexicales des mots. Ces propriétés influencent l'encodage perceptif initial et la rétention des mots parlés et doivent par conséquent jouer un rôle important dans les hypothèses théoriques concernant la façon dont le système nerveux associe les signaux de parole aux représentations linguistiques du lexique mental.

Keywords. Speech perception; perceptual normalization; long-term memory; talker variability; speaking rate; implicit memory; acoustic–phonetic variability; procedural memory; non-analytic perception; exemplar-based encoding; indexical properties of speech.

1. Introduction

My research in the early 1970's was directly motivated by Hiroya Fujisaki's proposal of the differential roles of auditory and phonetic memory codes in the perception of consonants and vowels (Fujisaki and Kawashima, 1969). The studies that I carried out at that time demonstrated that it was possible to account for categorical and non-categorical modes of perception in terms of coding and memory processes in short-term memory without recourse to the traditional theoretical accounts that were very popular at the time (Pisoni, 1973). These accounts of speech perception drew heavily on claims for a specialized perceptual mode for speech sounds that was distinct from other perceptual systems (Liberman et al., 1967).

Professor Fujisaki's efforts along with other results were largely responsible for integrating the study of speech perception with other closely related fields of cognitive psychology such as perception, memory and attention. By the mid 1970's, the field of speech perception became a legitimate topic for experimental psychologists to study (Pisoni, 1978). This was clearly an exciting time to be working in speech perception. Before these developments, speech perception was an exotic field representing the intersection of electrical engineering, speech science, linguistics, and traditional experimental psychology.

At the present time, the field of speech perception has evolved into an extremely active area of research with scientists from many different disciplines working on a common set of problems (Pisoni and Luce, 1987). Many of the current problems revolve around issues of representation and the role of coding and memory systems in spoken language processing, topics that Professor Fujisaki has written about in some detail over the years. The recent meetings of the ICSLP in Kobe and Banff demonstrate a convergence on a "core" set of basic research problems in the field of spoken language processing – problems that are inherently multi-disciplinary in nature. As many of us know from personal experiences, Professor Fujisaki was among the very first to recognize these common issues in his research and theoretical work over the years. The success of the two ICSLP meetings is due, in part, to his vision for a unified approach to the field of spoken language processing.

In this contribution, I am delighted to have the opportunity to summarize some recent work from my laboratory that deals with the role of long-term memory in speech perception and spoken word recognition. Much of our research over the last few years has turned to questions concerning perceptual learning and the retention of information in permanent long-term memory. This trend contrasts with the earlier work in the 1970's which was concerned almost entirely with short-term memory. We have also focused much of our current research on problems of spoken word recognition in contrast to earlier studies which were concerned with phoneme perception. We draw a distinction between phoneme perception and spoken word recognition. While phoneme perception is assumed to be a component of the word recognition process, the two are not equivalent. Word recognition entails access to phonological information stored in long-term memory, whereas phoneme perception relies almost exclusively on the recognition of acoustic cues contained in the speech signal.

Our interests are now directed at the interface between speech perception and spoken language comprehension which naturally has led us to problems of lexical access and the structure and organization of sound patterns in the mental lexicon (Pisoni et al., 1985). Findings from a variety of studies suggest that very fine details in the speech signal are preserved in the human memory system for relatively long periods of time (Goldinger, 1992). This information appears to be used in a variety of ways to facilitate perceptual encoding, retention and retrieval of information from memory. Many of our recent investigations have been concerned with assessing the effects of different sources of variability in speech perception (Sommers et al., 1992a; Nygaard et al., 1992a). The results of these studies have encouraged us to reassess our beliefs about several long-standing issues such as acoustic–phonetic invariance and the problems of perceptual normalization in speech perception (Pisoni, 1992a).

In the sections below, I will briefly summarize the results from several recent studies that deal with talker variability, speaking rate, and perceptual learning. These findings have raised a number of important new questions about the traditional dissociation between the linguistic and indexical properties of speech signals and the role that different sources of variability play in speech perception and spoken word recognition. For many years, linguists and phoneticians have considered attributes of the talker's voice – what Ladefoged refers to as the "personal" characteristics of speech – to be independent of the linguistic content of the talker's message (Ladefoged, 1975; Laver and Trudgill, 1979). The dissociation of these two parallel sources of information in speech may have served a useful function in the formal linguistic analysis of language when viewed as an idealized abstract system of symbols. However, the artificial dissociation has at the same time created some difficult problems for researchers who wish to gain a detailed understanding of how the nervous system encodes speech signals and represents them internally and how real speakers and listeners deal with the enormous amount of acoustic variability in speech.

2. Experiments on talker variability in speech perception

A series of novel experiments has been carried out to study the effects of different sources of variability on speech perception and spoken word recognition (Pisoni, 1990). Instead of reducing or eliminating variability in the stimulus materials, as most researchers had routinely done in the past, we specifically introduced variability from different talkers and different speaking rates to study their effects on perception (Pisoni, 1992b). Our research on talker variability began with the observations of Mullennix et al. (1989) who found that the intelligibility of isolated spoken words presented in noise was affected by the number of talkers that were used to generate the test words in the stimulus ensemble. In one condition, all the words in a test list were produced by a single talker; in another condition, the words were produced by 15 different talkers, including male and female voices. The results, which are shown in Figure 1, were very clear. Across three signal-to-noise ratios, identification performance was always better for words that were produced by a single talker than words produced by multiple talkers. Trial-to-trial variability in the speaker's voice apparently affects recognition performance. This pattern was observed for both high-density (i.e., confusable) and low-density (i.e., non-confusable) words. These findings replicated results originally found by Peters (1955) and Creelman

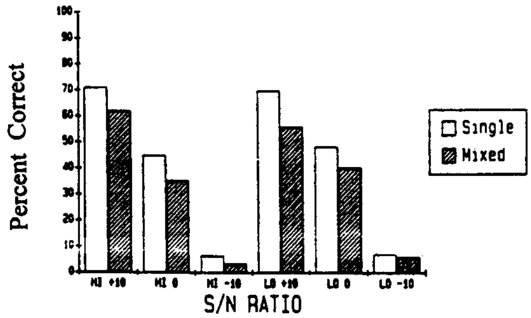


Fig. 1. Overall mean percent correct performance collapsed over subjects for single- and mixed-talker conditions as a function of high- and low-density words and signal-to-noise ratio (from (Mullennix et al., 1989)).

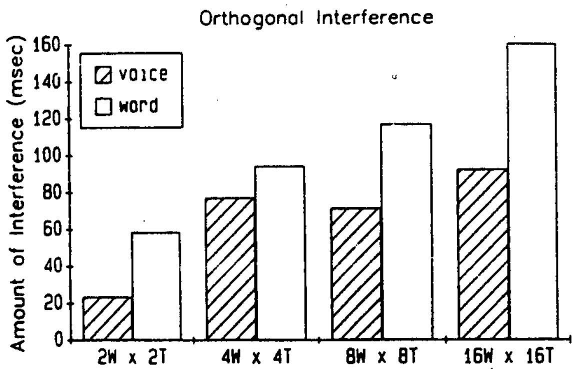


Fig. 2. The amount of orthogonal interference (in milliseconds) across all stimulus variability conditions as a function of word and voice dimensions (from (Mullennix and Pisoni, 1990)).

(1957) back in the 1950's and suggested to us that the perceptual system must engage in some form of "recalibration" each time a new voice is encountered during the set of test trials.

In a second experiment, we measured naming latencies to the same words presented in both test conditions (Mullennix et al., 1989). Table 1 provides a summary of the major results. We found that subjects were not only slower to name words from multiple-talker lists but they were also less accurate when their performance was compared to naming words from single-talker lists. Both sets of findings were surprising to us at the time because all the test words used in the experiment were highly intelligible when presented in the quiet. The intelligibility and naming data immediately raised a number of additional questions about how the various perceptual dimensions of the speech signal are processed by the human listener. At the time, we naturally assumed that the acoustic attributes used to perceive voice quality were independent of the linguistic properties of the signal. However, no one had ever tested this assumption directly.

Table 1
Mean response latency (ms) for correct responses for single- and mixed-talker conditions as a function of lexical density (from (Mullennix et al., 1989))

	Density	
	High	Low
Single talker	611.2	605.7
Mixed talker	677.2	679.4

In another series of experiments we used a speeded classification task to assess whether attributes of a talker's voice were perceived independently of the phonetic form of the words (Mullennix and Pisoni, 1990). Subjects were required to attend selectively to one stimulus dimension (i.e., voice) while simultaneously ignoring another stimulus dimension (i.e., phoneme). Figure 2 shows the main findings. Across all conditions, we found increases in interference from both dimensions when the subjects were required to attend selectively to only one of the stimulus dimensions. The pattern of results suggested that words and voices were processed as integral dimensions; the perception of one dimension (i.e., phoneme) affects classification of the other dimension (i.e., voice) and vice versa, and subjects cannot selectively ignore irrelevant variation on the non-attended dimension. If both perceptual dimensions were processed separately, as we originally assumed, we should have found little if any interference from the non-attended dimension which could be selectively ignored without affecting performance on the attended dimension. Not only did we find mutual interference suggesting that the two sets of dimensions, voice and phoneme, are perceived in a mutually dependent manner but we also found that the pattern of interference was asymmetrical. It was easier for subjects to ignore irrelevant variation in the phoneme dimension when their task was to classify the voice dimension than it was to ignore the voice dimension when they had to classify the phonemes.

The results from the perceptual experiments were surprising given our prior assumption that the indexical and linguistic properties of speech were perceived independently. To study this problem further, we carried out a series of memory experiments to assess the mental representation of speech in long-term memory. Experiments on serial recall of lists of spoken words by Martin et al. (1989) and Goldinger et al. (1991) demonstrated that specific details of a talker's voice are also encoded into long-term memory. Using a continuous recognition memory procedure, Palmeri et al. (1993) found that detailed episodic information about a talker's voice is also encoded in memory and is available for explicit judgments even when a great deal of competition from other voices is present in the test sequence. Palmeri et al.'s results are shown in Figure 3. The top panel shows the probability that an item was correctly recognized as a function of the number of talkers in the stimulus set. The bottom panel shows the probability of a correct recognition across different stimulus lags of intervening items. In both cases, the probability of correctly recognizing a word as "old" (filled circles) was greater if the word was repeated in the same voice than if it was repeated in a different voice of the same gender (open squares) or a different voice of a different gender (open triangles).

Finally, in another set of experiments, Goldinger (1992) found very strong evidence of implicit memory for attributes of a talker's voice which persists for a relatively long period of time after perceptual analysis has been completed. His results are shown in Figure 4. Goldinger also showed that the degree of perceptual similarity affects the magnitude of the repetition effect suggesting that the perceptual system encodes very detailed talker-specific information about spoken words in episodic memory.

Taken together, our findings on the effects of talker variability in perception and memory tasks provide support for the proposal that detailed perceptual information about a talker's voice is preserved in some type of perceptual representation system (PRS) (Schacter, 1990) and that these

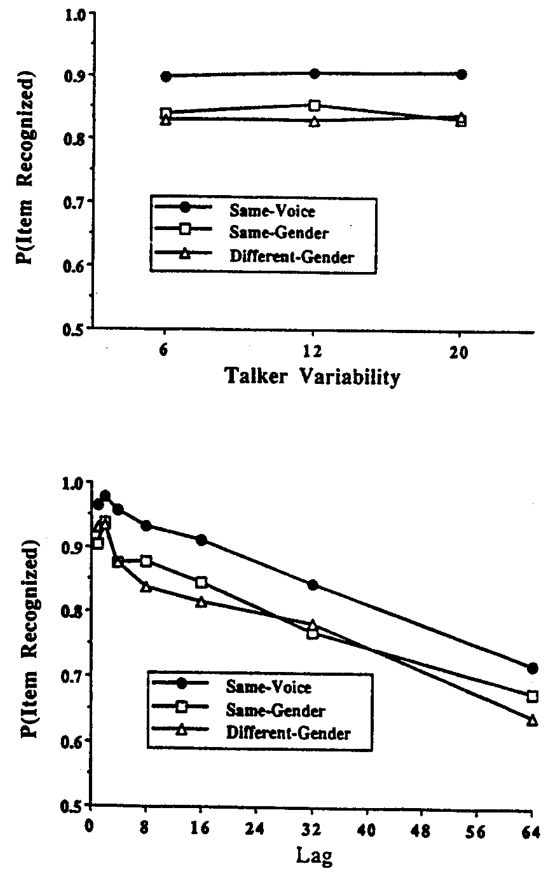


Fig. 3. Probability of correctly recognizing old items in a continuous recognition memory experiment. In both panels, recognition for same-voice repetitions is compared to recognition for different-voice/same-gender and different-voice/different-gender repetitions. The upper panel displays item recognition as a function of talker variability, collapsed across values of lag; the lower panel displays item recognition as a function of lag, collapsed across levels of talker variability (from (Palmeri et al., 1993)).

attributes are encoded into long-term memory. At the present time, it is not clear whether there is one composite representation in memory or whether these different sets of attributes are encoded in parallel in separate representations (Eich, 1982; Hintzman, 1986). It is also not clear whether spoken words are encoded and represented in memory as a sequence of abstract symbolic phoneme-like units along with much more detailed episodic information about specific instances and the processing operations used in

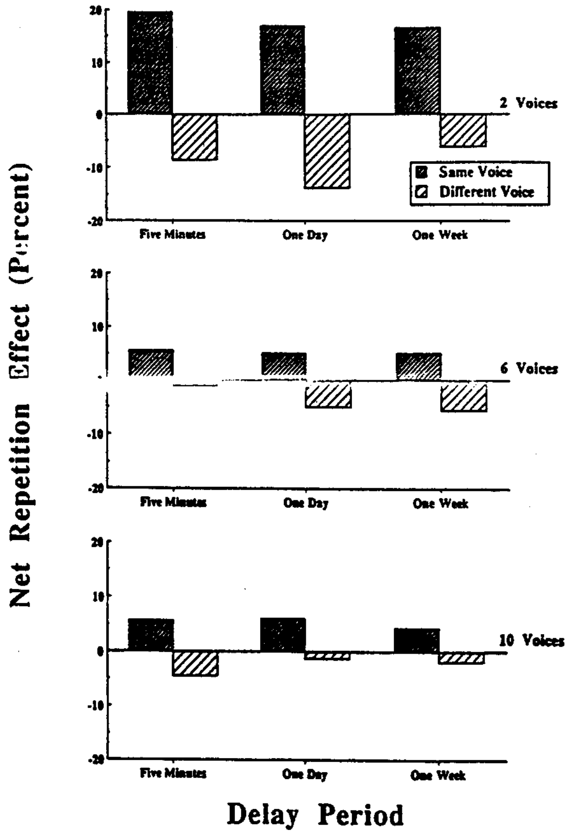


Fig. 4. Net repetition effects observed in perceptual identification as a function of delay between sessions and repetition voice (from (Goldinger, 1992)).

perceptual analysis. These are important questions for future research on spoken word recognition.

3. Experiments on the effects of speaking rate

Another new series of experiments has been carried out to examine the effects of speaking rate on perception and memory. These studies, which were designed to parallel the earlier experiments on talker variability, have also shown that the perceptual details associated with differences in speaking rate are not lost as a result of perceptual analysis. In one experiment, Sommers et al. (1992b) found that words produced at different speaking rates (i.e., fast, medium and slow) were identified more poorly than the same words produced at only one speaking rate. These results were compared to another condition in which differences in amplitude were varied randomly from trial to trial in the test sequences. In this case, identification performance was not affected by variability in overall level. The results from both conditions are shown in Figures 5 and 6.

Other experiments on serial recall have also been completed to examine the encoding and representation of speaking rate in memory. Nygaard et al. (1992b) found that subjects recall words from lists produced at a single speaking rate better than the same words produced at several different speaking rates. Interestingly, the differences appeared in the primacy portion of the serial position curve suggesting greater difficulty in the transfer of items into long-term memory. Differences in speaking rate, like those observed for talker variability in our earlier experiments, suggest that perceptual encoding and re-

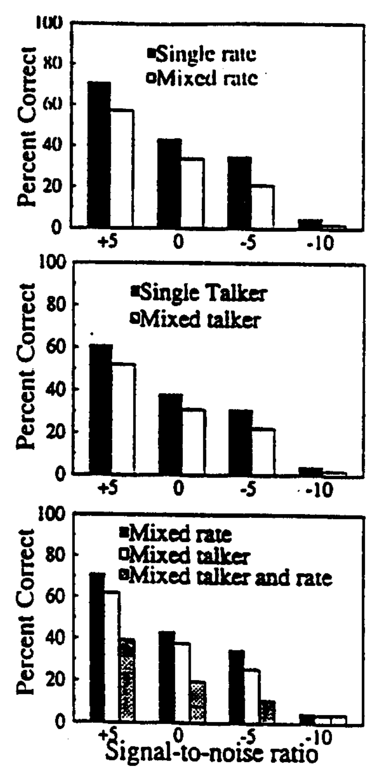


Fig. 5. Effects of talker, rate, and combined talker and rate variability on perceptual identification (from(Sommers et al., 1992b)).

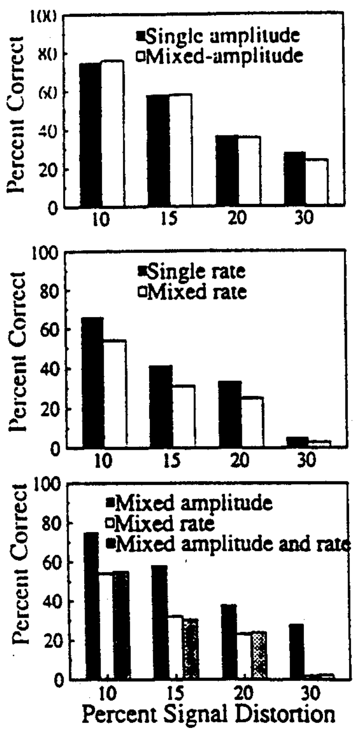


Fig. 6. Effects of amplitude, rate, and combined amplitude and rate variability on perceptual identification (from (Sommers et al., 1992)).

hearsal processes, which are typically thought to operate on only abstract symbolic representations, are also influenced by low-level perceptual sources of variability. If these sources of variability were somehow "filtered out" or normalized by the perceptual system at relatively early stages of analysis, differences in recall performance would not be expected in memory tasks like the ones used in these experiments.

Taken together with the earlier results on talker variability, the findings on speaking rate suggest that details of the early perceptual analysis of spoken words are not lost and apparently become an integral part of the mental representation of spoken words in memory. In fact, in some cases, increased stimulus variability in an experiment may actually help listeners to encode items into long-term memory (Goldinger et al., 1991; Nygaard et al., 1992b). Listeners encode speech signals in multiple ways along many perceptual dimensions and the memory system apparently preserves these perceptual details much more reliably than researchers have believed in the past.

4. Experiments on variability in perceptual learning

We have always maintained a strong interest in issues surrounding perceptual learning and development in speech perception (Aslin and Pisoni, 1980; Walley et al., 1981). One reason for this direction in our research is that much of the theorizing that has been done in speech perception has focused almost entirely on the mature adult with little concern for the processes of perceptual learning and developmental change. This has always seemed to be a peculiar state of affairs because it is now very well established that the linguistic environment plays an enormous role in shaping and modifying the speech perception abilities of infants and young children as they acquire their native language (Jusczyk, 1993). Theoretical accounts of speech perception should not only describe the perceptual abilities of the mature listener but they should also provide some principled explanations of how these abilities develop and how they are selectively modified by the language learning environment (Jusczyk, 1993; Studdert-Kennedy, 1980).

One of the questions that we have been interested in deals with the apparent difficulty that adult Japanese listeners have in discriminating English /r/ and /l/ (Logan et al., 1991; Lively et al., 1992, 1993; Strange and Dittmann, 1984). Is the failure to discriminate this contrast due to some permanent change in the perceptual abilities of native speakers of Japanese or are the basic sensory and perceptual mechanisms still intact and only temporarily modified by changes in selective attention and categorization? Many researchers working in the field have maintained the view that the effects of linguistic experience on speech perception are extremely difficult, if not impossible, to modify in a short period of time. The process of "re-learning" or "re-acquisition" of phonetic contrasts is generally assumed to be very difficult – it is slow, effortful and considerable variability has been observed among individuals in reacquiring sound contrasts that were not present in their native language (Strange and Dittmann, 1984).

We have carried out a series of laboratory training experiments to learn more about the difficulty Japanese listeners have in identifying English words containing /r/ and /l/ (Logan et al., 1991). In these studies we have taken some clues from the literature in cognitive psychology on the development of new perceptual categories and have designed our training procedures to capitalize on the important role that stimulus variability plays in perceptual learning (Posner and Keele, 1986). In the training phase of our experiments, we used a set of stimuli that contained a great deal of variability. The phonemes /r/ and /l/ appeared in English words in several different phonetic environments so that listeners would be exposed to different contextual variants of the same phoneme in different positions. In addition, we created a large database of words that were produced by several different talkers including both men and women in order to provide the listeners with exposure to a wide range of stimulus tokens.

A pretest–posttest design was used to assess the effects of the training procedures. Subjects were required to come to the laboratory for daily training sessions in which immediate feedback was provided after each trial. We trained a group of six Japanese listeners using a two-alternative forced-choice identification task. The stimulus materials consisted of minimal pairs of English words that contrasted /r/ and /l/ in five different phonetic environments.

On each training trial, subjects were presented with a minimal pair of words contrasting /r/ and /l/ on a CRT monitor. Subjects then heard one member of the pair and were asked to press a response button corresponding to the word they heard. If a listener made a correct response, the series of training trials continued. If a listener made an error, the minimal pair remained on the monitor and the stimulus word was repeated. In addition to the daily training sessions, subjects were also given a pretest and a posttest. At the end of the experiment, we also administered two additional tests of generalization. One test contained new words produced by one of the talkers used in training; the other test contained new words produced by a novel talker.

Identification accuracy improved significantly from the pretest to the posttest. Large and reliable effects of phonetic environment also were observed. Subjects were most accurate at identifying /r/ and /l/ in word final position. A significant interaction between the phonetic environment and pretest–posttest variables also was observed. Subjects improved more in initial consonant clusters and in intervocalic position than in word-initial and word-final positions.

The training results also showed that subjects' performance improved as a function of training. The largest gain came after one week of training. The gain in the other weeks was slightly smaller. Each of the six subjects showed improvement, although large individual differences in absolute levels of performance were observed.

The tests of generalization provided an additional way of assessing the effectiveness of the training procedures. Subjects were presented with new words spoken by a familiar talker and new words spoken by a novel talker. The /r/–/l/ contrast occurred in all five phonetic environments and listeners were required to perform the same categorization task. In our first training study, accuracy was marginally greater for words produced by the old talker compared to the new talker. However, in a replication experiment using 19 mono-lingual Japanese listeners, we found a highly significant difference in performance on the generalization tests (Lively et al., 1992). The results of the generalization tests demonstrate the high degree of context sensitivity present in learning to perceive these contrasts: Listeners were sensitive to the voice of the talker producing the tokens as well as the phonetic environment in which the contrasts occurred. Thus, stimulus variability is useful in perceptual learning of complex multidimensional categories like speech because it serves to make the mental representations extremely robust over different acoustic transformations such as talker, phonetic environment and speaking rate. In a high variability training procedure, like the one used by Logan et al., listeners are not able to focus their attention on only one set of criterial cues to learn the category structure for the phonemes /r/ and /l/. Listeners have to acquire detailed knowledge about different sources of variability in order to be able to generalize to new words and new talkers.

We have also been interested in another kind of perceptual learning, the tuning or adaptation that occurs when a listener becomes familiar with the voice of a specific talker (Nygaard et al., in press). This particular kind of perceptual learning has not received very much attention in the past despite the obvious relevance to problems of speaker normalization, acoustic–phonetic invariance and the potential application to automatic speech recognition and speaker identification (Kakehi, 1992; Fowler, in press). Our search of the research literature on talker adaptation revealed only a small number of studies on this topic and all of them appeared in obscure technical reports from the mid 1950's. Thus, we decided to carry out a perceptual learning experiment in our own laboratory.

To determine how familiarity with a talker's voice affects the perception of spoken words, we had listeners learn to explicitly identify a set of unfamiliar voices over a nine day period using common names (i.e., Bill, Joe, Sue, Mary). After the subjects learned to recognize the voices, we presented them with a set of novel words mixed in noise at several signal-to-noise ratios; half the listeners heard the words produced by talkers that they were previously trained on and half the listeners heard the words produced by new talkers that they had not been exposed to previously. In this phase of the experiment, which was designed to measure speech intelligibility, subjects were required to identify the words rather than recognize the voices as they had done in the earlier phase of the experiment.

The results of the intelligibility experiment are shown in Figure 7 for two groups of subjects. We found that identification performance for the trained group was reliably better than the control group at each of the signal-to-noise ratios tested. The subjects who had heard novel words produced by familiar voices were able to recognize words in noise more accurately than subjects who received the same novel words produced by unfamiliar voices. Two other groups of subjects were also run in the intelligibility experiment as controls; however, these subjects did not receive any training and were therefore not exposed to any of the voices prior to hearing the same set of words in noise. One control group received the set of words presented to the trained experimental group; the other control group received the words that were presented to the trained control subjects. The performance of these two control groups was not only same but was equivalent to the intelligibility scores obtained by the trained control group. Only subjects in the experimental group who were explicitly trained on the voices showed an advantage in recognizing novel words produced by familiar talkers.

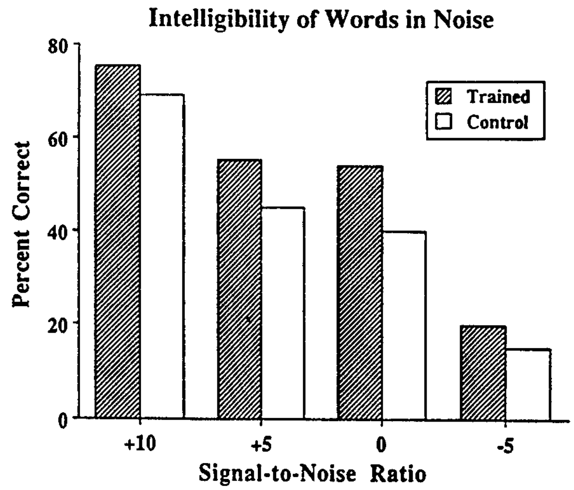


Fig. 7. Mean intelligibility of words mixed in noise for trained and control subjects. Percent correct word recognition is plotted at each signal-to-noise ratio (from (Nygaard et al., 1992b)).

The findings from this perceptual learning experiment demonstrate that exposure to a talker's voice facilitates subsequent perceptual processing of novel words produced by a familiar talker. Thus, speech perception and spoken word recognition draw on highly specific perceptual knowledge about a talker's voice that was obtained in an entirely different experimental task – explicit voice recognition as compared to a speech intelligibility test in which novel words were mixed in noise and subjects identified the items explicitly from an open response set.

What kind of perceptual knowledge does a listener acquire when he listens to a speaker's voice and is required to carry out an explicit name recognition task like our subjects did in this experiment? One possibility is that the procedures or perceptual operations (Kolers, 1973) used to recognize the voices are retained in some type of "procedural memory" and these routines are invoked again when the same voice is encountered in a subsequent intelligibility test. This kind of procedural knowledge might increase the efficiency of the perceptual analysis for novel words produced by familiar talkers because detailed analysis of the speaker's voice would not have to be carried out again. Another possibility is that specific instances – perceptual episodes or exemplars of each talker's voice – are stored in memory and then later retrieved during the process of word recognition when new tokens from a familiar talker are encountered (Jacoby and Brooks, 1984).

Whatever the exact nature of this information or knowledge turns out to be, the important point here is that prior exposure to a talker's voice facilitates subsequent recognition of novel words produced by the same talkers. Such findings demonstrate a form of implicit memory for a talker's voice that is distinct from the retention of the individual items used and the specific task that was employed to familiarize the listeners with the voices (Schacter, 1992; Roediger, 1990). These findings provide additional support for the view that the internal representation of spoken words encompasses both a phonetic description of the utterance, as well as information about the structural description of the source characteristics of the specific talker. Thus, speech perception appears to be carried out in a "talker-contingent" manner; indexical and linguistic properties of the speech signal are apparently closely interrelated and are not dissociated in perceptual analysis as many researchers previously thought. We believe these talker-contingent effects may provide a new way to deal with some of the old problems in speech perception that have been so difficult to resolve in the past.

5. Abstractionist versus episodic approaches to speech perception

The results we have obtained over the last few years raise a number of important questions about the theoretical assumptions or metatheory of speech perception which has been shared for many years by almost all researchers working in the field (Pisoni and Luce, 1986). Within cognitive psychology, the traditional view of speech perception can be considered among the best examples of what have been called "abstractionist" approaches to the problems of categorization and memory (Jacoby and Brooks, 1984). Units of perceptual analysis in speech were assumed to be equivalent to the abstract idealized categories proposed by linguists in their formal analyses of language structure and function. The goal of speech perception studies was to find the physical invariants in the speech signal that mapped onto the phonetic categories of speech (Studdert-Kennedy, 1976). Emphasis was directed at separating stable, relevant features from the highly variable, irrelevant features of the signal. An important assumption of this traditional approach to perception and cognition was the process of abstraction and the reduction of information in the signal to a more efficient and economical symbolic code (Posner, 1969; Neisser, 1976). Unfortunately, it became apparent very early on in speech perception research that idealized linguistic units, such as phonemes or phoneme-like units, were highly dependent on phonetic context and moreover that a wide variety of factors influenced their physical realization in the speech signal (Stevens, 1971; Klatt, 1986). Nevertheless, the search for acoustic invariance has continued in one way or another and still remains a central problem in the field today.

Recently, a number of studies on categorization and memory in cognitive psychology have provided evidence for the encoding and retention of episodic information and the details of perceptual analysis (Jacoby and Brooks, 1984; Brooks, 1978; Tulving and Schacter, 1990; Schacter, 1990). According to this approach, stimulus variability is considered to be "lawful" and informative to perceptual analysis (Elman and McClellan, 1986). Memory involves encoding specific instances, as well as the processing operations used in recognition (Kolers, 1973, 1976). The major emphasis of this view is on particulars, rather than abstract generalizations or symbolic coding of the stimulus input into idealized categories. Thus, the problems of variability and invariance in speech perception can be approached in a different way by non-analytic or instance-based accounts of perception and memory with the emphasis on encoding of exemplars and specific instances of the stimulus environment rather than the search for physical invariants for abstract symbolic categories.

We believe that the findings from studies on nonanalytic cognition can be generalized to theoretical questions about the nature of perception and memory for speech signals and to assumptions about abstractionist representations based on formal linguistic analyses. When the criteria used for postulating episodic or non-analytic representations are examined carefully, it immediately becomes clear that speech signals display a number of distinctive properties that make them especially good candidates for this approach to perception and memory (Jacoby and Brooks, 1984; Brooks, 1978). These criteria which are summarized below can be applied directly to speech perception and spoken language processing.

5.1. High stimulus variability

Speech signals display a great deal of variability primarily because of factors that influence the production of spoken language. Among these are within- and between-talker variability, changes in speaking rate and dialect, differences in social contexts, syntactic, semantic and pragmatic effects, as well as a wide variety of effects due to the ambient environment such as background noise, reverberation and microphone characteristics (Klatt, 1986). These diverse sources of variability consistently produce large changes in the acoustic–phonetic properties of speech and they need to be accommodated in theoretical accounts of speech perception.

5.2. Complex category relations

The use of phonemes as perceptual categories in speech perception entails a set of complex assumptions about category membership which are based on formal linguistic criteria involving principles such as complementary distribution, free variation and phonetic similarity. The relationship between allophones and phonemes acknowledges explicitly the context-sensitive nature of the category relations that are used to define classes of speech sounds that function in similar ways in different phonetic environments. In addition, there is evidence for "trading relations" among cues to particular phonetic contrasts in speech. Acoustically different cues to the same contrast interact as a function of context.

5.3. Incomplete information

Spoken language is a highly redundant symbolic system which has evolved to maximize transmission of information. In the case of speech perception, research has demonstrated the existence of multiple speech cues for almost every phonetic contrast. While these speech cues are, for the most part, highly context-dependent, they also provide partial information that can facilitate comprehension of the intended message when the signal is degraded. This feature of speech perception permits high rates of information transmission even under poor listening conditions.

5.4. High analytic difficulty

Speech sounds are inherently multidimensional in nature. They encode a large number of quasi-independent articulatory attributes that are mapped on to the phonological categories of a specific language. Because of the complexity of speech categories and the high acoustic–phonetic variability, the category structure of speech is not amenable to simple hypothesis testing. As a consequence, it has been extremely difficult to formalize a set of explicit rules that can successfully map speech cues onto a set of idealized phoneme categories. Phoneme categories are also highly automatized. The category structure of a language is learned in a tacit and incidental way by young children. Moreover, because the criterial dimensional structures of speech are not typically available to consciousness, it is difficult to make many aspects of speech perception explicit to either children, adults, or machines.

5.5. Three domains of speech

Among category systems, speech appears to be unusual in several respects because of the mapping between production and perception. Speech exists simultaneously in three very different domains: the acoustic domain, the articulatory domain and the perceptual domain. While the relations among these three domains is complex, they are not arbitrary because the sound contrasts used in a language function within a common linguistic signaling system that is assumed to encompass both production and perception. Thus, the phonetic distinctions generated in speech production by the vocal tract are precisely those same acoustic differences that are important in perceptual analysis (Stevens, 1972). Any theoretical account of speech perception must also take into consideration aspects of speech production and acoustics. The perceptual spaces mapped out in speech production have to be very closely correlated with the same ones used in speech perception.

In learning the sound system of a language, the child must not only develop abilities to discriminate and identify sounds, but he/she must also be able to control the motor mechanisms used in articulation to generate precisely the same phonetic contrasts in speech production that he/she has become attuned to in perception. One reason that the developing perceptual system might preserve very fine phonetic details as well as characteristics of the talker's voice would be to allow a young child to accurately imitate and reproduce speech patterns heard in the surrounding language learning environment (Studdert-Kennedy, 1983). This skill would provide the child with an enormous benefit in acquiring the phonology of the local dialect from speakers he/she is exposed to early in life.

6. Discussion

It has become common over the last 25 years to argue that speech perception is a highly unique process that requires specialized neural processing mechanisms to carry out perceptual analysis (Liberman et al., 1967). These theoretical accounts of speech perception have typically emphasized the differences in perception between speech and other perceptual processes. Relatively few researchers working in the field of speech perception have tried to identify commonalities among other perceptual systems and draw parallels with speech. Our recent findings on the encoding of different sources of variability in speech and the role of long-term memory for specific instances are compatible with a rapidly growing body of research in cognitive psychology on implicit memory phenomena and non-analytic modes of processing (Jacoby and Brooks, 1984; Brooks, 1978).

Traditional memory research has been concerned with "explicit memory" in which the subject is required to consciously access and manipulate recently presented information from memory using "direct tests" such as recall or recognition. This line of memory research has a long history in experimental psychology and it is an area that most speech researchers are familiar with. In contrast, the recent literature on "implicit memory" phenomena has provided new evidence for unconscious aspects of perception, memory and cognition (Schacter, 1992; Roediger, 1990). Implicit memory refers to a form of memory that was acquired during a specific instance or episode and it is typically measured by "indirect tests" such as stem or fragment completion, priming or changes in perceptual identification performance. In these types of memory tests, subjects are not required to consciously recollect previously acquired information. In fact, in many cases, especially in processing spoken language, subjects may be unable to access the information deliberately or even bring it to consciousness (Studdert-Kennedy, 1974).

Studies of implicit memory have uncovered important new information about the effects of prior experience on perception and memory. In addition to traditional abstractionist modes of cognition which tend to emphasize symbolic coding of the stimulus input, numerous recent experiments have provided evidence for a parallel non-analytic memory system that preserves specific instances of stimulation as perceptual episodes or exemplars which are also stored in memory. These perceptual episodes have been shown to affect later processing activities. We believe that it is this implicit perceptual memory system that encodes the indexical information in speech about talker's gender, dialect and speaking rate. And, we believe that it is this memory system that encodes and preserves the perceptual operations or procedural knowledge that listeners acquire about specific voices that facilitates later recognition of novel words by familiar speakers.

Our findings demonstrating that spoken word recognition is talker-contingent and that familiar voices are encoded differently than novel voices raises a new set of questions concerning the long-standing dissociation between the linguistic properties of speech – the features, phonemes and words used to convey the linguistic message – and the indexical properties of speech – those personal or paralinguistic attributes of the speech signal which provide the listener with information about the form of the message – the speaker's gender, dialect, social class, and emotional state, among other things. In the past, these two sources of information were separated for purposes of linguistic analysis of the message. The present set of findings suggests this may have been an incorrect assumption for speech perception.

Relative to the research carried out on the linguistic properties of speech, which has a history dating back to the late 1940's, much less is known about perception of the acoustic correlates of the indexical or paralinguistic functions of speech (Ladefoged, 1975; Laver and Trudgill, 1979). While there have been a number of recent studies on explicit voice recognition and identification by human listeners (Papcun et al., 1989), very little research has been carried out on problems surrounding the "implicit" or "unconscious" encoding of attributes of voices and how this form of memory might affect the recognition process associated with the linguistic attributes of spoken words (Nygaard et al., in press). A question that naturally arises in this context is whether or not familiar voices are processed differently than unfamiliar or novel voices. Perhaps familiar voices are simply recognized more efficiently than novel voices and are perceived in fundamentally the same way by the same neural mechanisms as unfamiliar voices. The available evidence in the literature has shown, however, that familiar and unfamiliar voices are processed differentially by the two hemispheres of the brain and that selective impairment resulting from brain damage can affect the perception of familiar and novel voices in very different ways (Kreiman and VanLancker, 1988; VanLancker et al., 1988, 1989).

Most researchers working in speech perception adopted a common set of assumptions about the units of linguistic analysis and the goals of speech perception. The primary objective was to extract the speaker's message from the acoustic waveform without regard to the source (Studdert-Kennedy, 1974). The present set of findings suggests that while the dissociation between indexical and linguistic properties of speech may have been a useful dichotomy for theoretical linguists who approach language as a highly abstract formalized symbolic system, the same set of assumptions may no longer be useful for speech scientists who are interested in describing and modeling how the human nervous system encodes speech signals into representations in long-term memory.

Our recent findings on variability suggest that fine phonetic details about the form and structure of the signal are not lost as a consequence of perceptual analysis as widely assumed by researchers years ago. Attributes of the talker's voice are also not lost or normalized away, at least not immediately after perceptual analysis has been completed. In contrast to the theoretical views that were very popular a few years ago, the present findings have raised some new questions about how researchers have approached the problems of variability, invariance and perceptual normalization in the past. For example, there is now sufficient evidence from perceptual experimentation to suggest that the fundamental perceptual categories of speech – phonemes and phoneme-like units – are probably not as rigidly fixed or well-defined physically as theorists once believed. These perceptual categories appear to be highly variable and their physical attributes have been shown to be strongly affected by a wide variety of contextual factors (Klatt, 1979). It seems very unlikely after some 45 years of research on speech that very simple physical invariants for phonemes will be uncovered from analysis of the speech signal. If invariants are uncovered they will probably be very complex time-varying cues that are highly context-dependent.

Many of the theoretical views that speech researchers have held about language were motivated by linguistic considerations of speech as an idealized symbolic system essentially free from physical variability. Indeed, variability in speech was considered by many researchers to be a source of "noise" – an undesirable set of perturbations on what was otherwise supposed to be an idealized sequence of abstract symbols arrayed linearly in time. Unfortunately, it has taken a long time for speech researchers to realize that variability is an inherent characteristic of all biological systems including speech. Rather than view variability as noise, some theorists have recognized that variability might actually be useful and informative to human listeners who are able to encode speech signals in variety of different ways depending upon the circumstances and demands of the listening task (Elman and McClellan, 1986). The recent proposals in the human memory literature for multiple memory systems suggest that the internal representation of speech is probably much more detailed and more elaborate than previously believed from simply an abstractionist linguistic point of view. The traditional views about features, phonemes and acoustic–phonetic invariance are simply no longer adequate to accommodate the new findings that have been uncovered concerning context effects and variability in speech perception and spoken word recognition. In the future, it may be very useful to explore the parallels between similar perceptual systems such as face recognition and voice recognition. There is, in fact, some reason to suspect that parallel neural mechanisms may be employed in each case despite the obvious differences in modalities.

7. Conclusions

The results summarized in this paper on the role of variability in speech perception are compatible with non-analytic or instance-based views of cognition which emphasize the episodic encoding of specific details of the stimulus environment. Our studies on talker and rate variability and our new experiments on perceptual learning of novel phonetic contrasts and novel voices have provided important information about speech perception and spoken word recognition and have served to raise a set of new questions for future research. In this section, I simply list the major conclusions and hope these will encourage others to look at some of the long-standing problems in our field in a different way in the future.

First, our findings raise questions about previous views of the mental representation of speech. In particular, we have found that very detailed information about the source characteristics of a talker's voice is encoded into long-term memory. Whatever the internal representation of speech turns out to be, it is clear that it is not isomorphic with the linguist's description of speech as an abstract idealized sequence of segments. Mental representations of speech are much more detailed and more elaborate and they contain several sources of information about the talker's voice; perhaps these representations retain a perceptual record of the processing operations used to recognize the input patterns or maybe they reflect some other set of talker-specific attributes that permit a listener to explicitly recognize the voice of a familiar talker when asked to do so directly.

Second, our findings suggest a different approach to the problem of acoustic–phonetic variability in speech perception. Variability is not a source of noise; it is lawful and provides potentially useful information about characteristics of the talker's voice and speaking rate as well as the phonetic context. These sources of information may be accessed when a listener hears novel words or sentences produced by a familiar talker. Variability may provide important talker-specific information that affects encoding fluency and processing efficiency in a variety of tasks.

Third, our findings provide additional evidence that speech categories are highly sensitive to context and that some details of the input signal are not lost or filtered out as a consequence of perceptual analysis. These results are consistent with recent proposals for the existence of multiple memory systems and the role of perceptual representation systems (PRS) in memory and learning. The present findings also suggest a somewhat different view of the process of perceptual normalization which has generally focused on the processes of abstraction and stimulus reduction in categorization of speech sounds.

Finally, the results described here suggest several directions for new models of speech perception and spoken word recognition that are motivated by a different set of criteria than traditional abstractionist approaches to perception and memory. Exemplar-based or episodic models of categorization provide a viable theoretical alternative to the problems of invariance, variability and perceptual normalization that have been difficult to resolve with current models of speech perception that were inspired by formal linguistic analyses of language. We believe that many of the current theoretical problems in the field can be approached in quite different ways when viewed within the general framework of non-analytic or instance-based models of cognition which have alternative methods of dealing with variability, context effects and perceptual learning which have been the hallmarks of human speech perception.

Acknowledgments

This research was supported by NIDCD Research Grant DC-00111-17, Indiana University in Bloomington. I thank Steve Goldinger, Scott Lively, Lynne Nygaard, Mitchell Sommers and Thomas Palmeri for their help and collaboration in various phases of this research program.

References

R.N. Aslin and D.B. Pisoni (1980), "Some developmental processes in speech perception", in *Child Phonology: Perception and Production*, ed. by G. Yeni-Komshian, J.F. Kavanagh and C.A. Ferguson (Academic Press, New York), pp. 67–96.

L.R. Brooks (1978), "Nonanalytic concept formation and memory for instances", in *Cognition and Categorization*, ed. by E. Rosch and B. Lloyd (Erlbaum, Hillsdale, NJ), pp. 169–211.

C.D. Creelman (1957), "Case of the unknown talker", *J. Acoust. Soc. Amer.*, Vol. 29, p. 655.

J.E. Eich (1982), "A composite holographic associative memory model", *Psychol. Rev.*, Vol. 89, pp. 627–661.

J.L. Elman and J.L. McClellan (1986), "Exploiting lawful variability in the speech wave", in *Invariance and Variability in Speech Processes*, ed. by J.S. Perkell and D.H. Klatt (Erlbaum, Hillsdale, NJ), pp. 360–385.

C.A. Fowler (in press), "Listener–talker attunements in speech", in *Human Development and Communication Sciences*, edited by T. Tighe, B. Moore and J. Santroch (Erlbaum, Hillsdale, NJ).

H. Fujisaki and T. Kawashima (1969), "On the modes and mechanisms of speech perception", *Annual Report of the Engineering Research Institute*, Vol. 28, Faculty of Engineering, University of Tokyo, Tokyo, pp. 67–73.

S.D. Goldinger (1992), "Words and voices: Implicit and explicit memory for spoken words", *Research on Speech Perception Technical Report No. 7* (Indiana University, Bloomington, IN).

S.D. Goldinger, D.B. Pisoni and J.S. Logan (1991), "On the locus of talker variability effects in recall of spoken word lists", *J. Exp. Psychol.: Learning, Memory, and Cognition*, Vol. 17, No. 1, pp. 152–162.

D.L. Hintzman (1986), "Schema abstraction in a multiple-trace memory model", *Psychol. Rev.*, Vol. 93, pp. 411–423.

L.L. Jacoby and L.R. Brooks (1984), "Nonanalytic cognition: Memory, perception, and concept learning", in *The Psychology of Learning and Motivation*, ed. by G. Bower (Academic Press, New York), pp. 1–47.

P.W. Jusczyk (1993), "Infant speech perception and the development of the mental lexicon", in *The Transition from Speech Sounds to Spoken Words: The Development of Speech Perception*, ed. by H.C. Nusbaum and J.C. Goodman (MIT Press, Cambridge, MA).

K. Kakehi (1992), "Adaptability to differences between talkers in Japanese monosyllabic perception", in *Speech Perception, Production and Linguistic Structure*, ed. by Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (IOS Press, Tokyo), pp. 135–142.

D.H. Klatt (1979), "Speech perception: A model of acoustic–phonetic analysis and lexical access", *J. Phonet.*, Vol. 7, pp. 279–312.

D.H. Klatt (1986), "The problem of variability in speech recognition and in models of speech perception", in *Invariance and Variability in Speech Processes*, ed. by J.S. Perkell and D.H. Klatt (Erlbaum, Hillsdale, NJ), pp. 300–324.

P.A. Kolers (1973), "Remembering operations", *Memory & Cognition*, Vol. 1, pp. 347–355.

P.A. Kolers (1976), "Pattern analyzing memory", *Science*, Vol. 191, pp. 1280–1281.

J. Kreiman and D. VanLancker (1988), "Hemispheric specialization for voice recognition: Evidence from dichotic listening", *Brain & Language*, Vol. 34, pp. 246–252.

P. Ladefoged (1975), *A Course in Phonetics* (Harcourt Brace Jovanovich, New York).

J. Laver and P. Trudgill (1979), "Phonetic and linguistic markers in speech", in *Social Markers in Speech*, ed. by K.R. Scherer and H. Giles (Cambridge University Press, Cambridge), pp. 1–31.

A.M. Liberman, F.S. Cooper, D.P. Shankweiler and M. Studdert-Kennedy (1967), "Perception of the speech code", *Psychol. Rev.*, Vol. 74, pp. 431–461.

S.E. Lively, D.B. Pisoni and J.S. Logan (1992), "Some effects of training Japanese listeners to identify English /r/ and /l/", in *Speech Perception, Production and Linguistic Structure*, ed. by Y. Tohkura (Ohmsha, Tokyo), pp. 175–196.

S.E. Lively, D.B. Pisoni, R.A. Yamada, Y. Tohkura and T. Yamada (1992), Training Japanese listeners to identify English |r| and |l|: III. Long-term retention of new phonetic categories, Research on Speech Perception, Progess Report No. 18, Indiana University, pp. 185–216.

S.E. Lively, J.S. Logan and D.B. Pisoni (1993), "Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories", *J. Acoust. Soc. Amer.*

J.S. Logan, S.E. Lively and D.B. Pisoni (1991), "Training Japanese listeners to identify English /r/ and /l/: A first report", *J. Acoust. Soc. Amer.*, Vol. 89, No. 2, pp. 874–886.

C.S. Martin, J.W. Mullennix, D.B. Pisoni and W.V. Summers (1989), "Effects of talker variability on recall of spoken word lists", *J. Exp. Psychol.: Learning, Memory and Cognition*, Vol. 15, pp. 676–684.

J.W. Mullennix and D.B. Pisoni (1990), "Stimulus variability and processing dependencies in speech perception", *Perception & Psychophysics*, Vol. 47, No. 4, pp. 379–390.

J.W. Mullennix, D.B. Pisoni and C.S. Martin (1989), "Some effects of talker variability on spoken word recognition", *J. Acoust. Soc. Amer.*, Vol. 85, pp. 365–378.

U. Neisser (1976), *Cognitive Psychology* (Appleton–Century–Crofts, New York).

L.C. Nygaard, M.S. Sommers and D.B. Pisoni (1992a), "Effects of speaking rate and talker variability on the representation of spoken words in memory", *Proc. 1992 Internat. Conf. Spoken Language Processing, Banff, Canada*, pp. 209–212.

L.C. Nygaard, M.S. Sommers and D.B. Pisoni (1992b), "Effects of speaking rate and talker variability on the recall of spoken words", *J. Acoust. Soc. Amer.*, Vol. 91, No. 4, p. 2340.

L.C. Nygaard, M.S. Sommers and D.B. Pisoni (in press), "Speech perception as a talker-contingent process", *Psychol. Sci.*

T.J. Palmeri, S.D. Goldinger and D.B. Pisoni (1993), "Episodic encoding of voice attributes and recognition memory for spoken words", *J. Exp. Psychol.: Learning, Memory and Cognition*, Vol. 19, No. 2, pp. 1–20.

G. Papcun, J. Kreiman and A. Davis (1989). "Long-term memory for unfamiliar voices", *J. Acoust. Soc. Amer.*, Vol. 85, pp. 913–925.

R.W. Peters (1955), The relative intelligibility of single-voice and multiple-voice messages under various conditions of noise, Joint Project Report No. 56, U.S. Naval School of Aviation Medicine, Pensacola, FL, pp. 1–9.

D.B. Pisoni (1973), "Auditory and phonetic memory codes in the discrimination of consonants and vowels", *Perception & Psychophysics*, Vol. 13, pp. 253–260.

D.B. Pisoni (1978), "Speech perception", in *Handbook of Learning and Cognitive Processes*, ed. by W.K. Estes (Erlbaum, Hillsdale, NJ), Vol. 6, pp. 167–233.

D.B. Pisoni (1990), "Effects of talker variability on speech perception: Implications for current research and theory", *Proc. 1990 Internat. Conf. Spoken Language Processing, Kobe, Japan*, pp. 1399–1407.

D.B. Pisoni (1992a), "Some comments on invariance, variability and perceptual normalization in speech perception", *Proc. 1992 Internat. Conf. Spoken Language Processing, Banff, Canada*, pp. 587–590.

D.B. Pisoni (1992b), "Some comments on talker normalization in speech perception", in *Speech Perception, Production and Linguistic Structure*, ed. by Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (IOS Press, Tokyo), pp. 143–151.

D.B. Pisoni and P.A. Luce (1986), "Speech perception: Research, theory, and the principal issues", in *Pattern Recognition by Humans and Machines*, ed. by E.C. Schwab and H.C. Nusbaum (Academic Press, New York), pp. 1–50.

D.B. Pisoni and P.A. Luce (1987), "Acoustic–phonetic representations in word recognition", *Cognition*, Vol. 25, pp. 21–52.

D.B. Pisoni, H.C. Nusbaum, P.A. Luce and L.M. Slowiaczek (1985), "Speech perception, word recognition and the structure of the lexicon", *Speech Communication*, Vol. 4, Nos. 1–3, pp. 75–95.

M.I. Posner (1969), "Abstraction and the process of recognition", in *The Psychology of Learning and Motivation: Advances in Learning and Motivation*, ed. by J.T. Spence and G.H. Bower (Academic Press, New York), pp. 43–100.

M. Posner and S. Keele (1986), "On the genesis of abstract ideas", *J. Exp. Psychol.*, Vol. 77, pp. 353–363.

H.L. Roediger (1990), "Implicit memory: Retention without remembering", *Amer. Psychol.*, Vol. 45, No. 9, pp. 1043–1056.

D.L. Schacter (1990), "Perceptual representation systems and implicit memory: Toward a resolution of the multiple memory systems debate", in *Development and Neural Basis of Higher Cognitive Function*, edited by A. Diamond, *Ann. New York Acad. Sci.*, Vol. 608, pp. 543–571.

D.L. Schacter (1992), "Understanding implicit memory: A cognitive neuroscience approach", *Amer. Psychol.*, Vol. 47, No. 4, pp. 559–569.

M.S. Sommers, L.C. Nygaard and D.B. Pisoni (1992a), "Stimulus variability and the perception of spoken words: Effects of variations in speaking rate and overall amplitude", *Proc. 1992 Int. Conf. Spoken Language Processing, Banff, Canada,* pp. 217–220.

M.S. Sommers, L.C. Nygaard and D.B. Pisoni (1992b), "The effects of speaking rate and amplitude variability on perceptual identification", *J. Acoust. Soc. Amer.*, Vol. 91, No. 4, p. 2340.

K.N. Stevens (1971), "Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds", *Proc. 7th Int. Cong. Phonetic Sciences* (Mouton, The Hague), pp. 206–232.

K.N. Stevens (1972), "The quantal nature of speech: Evidence from articulatory acoustic data", in *Human Communica-*

tion: A Unified View, ed. by E.E. David, Jr. and P.B. Denes (McGraw-Hill, New York), pp. 51–66.

W. Strange and S. Dittmann (1984), "The effects of discrimination training on the perception of /r–l/ by Japanese adults learning English". *Perception & Psychophysics*, Vol. 36, No. 2, pp. 131–145.

M. Studdert-Kennedy (1974), "The perception of speech", in *Current Trends in Linguistics*, ed. by T.A. Sebeok (Mouton, The Hague), pp. 2349–2385.

M. Studdert-Kennedy (1976), "Speech perception", in *Contemporary Issues in Experimental Phonetics*, ed. by N.J. Lass (Academic Press, New York), pp. 243–293.

M. Studdert-Kennedy (1980), "Speech perception", *Language and Speech*, Vol. 23, pp. 45–66.

M. Studdert-Kennedy (1983), "On learning to speak", *Hum. Neurobiol.*, Vol. 2, pp. 191–195.

E. Tulving and D.L. Schacter (1990), "Priming and human memory systems", *Science*, Vol. 247, pp. 301–306.

D.R. VanLancker, J.L. Cummings, J. Kreiman and B. Dobkin (1988), "Phonagnosia: A dissociation between familiar and unfamiliar voices", *Cortex*, Vol. 24, pp. 195–209.

D.R. VanLancker, J. Kreiman and J. Cummings (1989), "Voice perception deficits: Neuroanatomical correlates of phonagnosia", *J. Clin. Exp. Neuropsychol.*, Vol. 11, No. 5, pp. 665–674.

A.C. Walley, D.B. Pisoni and R.N. Aslin (1981), "The role of early experience in the development of speech perception", in *The Development of Perception: Psychobiological Perspectives*, ed. by R.N. Aslin, J. Alberts and M.R. Petersen (Academic Press, New York), pp. 2119–255.

☆U.S. GOVERNMENT PRINTING OFFICE: 1996 - 760-082/20067