INVESTIGATING PILOT PERFORMANCE USING MIXED-MODALITY SIMULATED DATA LINK

by

Jeff A. Lancaster

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Industrial and Systems Engineering

Advisory Committee:

Dr. John G. Casali, Chairman
Dr. Gary S. Robinson
Dr. Brian M. Kleiner
Dr. Antonio A. Trani
Dr. Thurmon E. Lockhart

April 9, 2004
Blacksburg, Virginia

Keywords: Pilot Performance, Data Link, Workload, Situation Awareness, Mixed-Modality

INVESTIGATING PILOT PERFORMANCE USING MIXED-MODALITY SIMULATED DATA LINK

by

Jeff A. Lancaster

Chairman: Dr. John G. Casali

Industrial and Systems Engineering

(ABSTRACT)

Empirical studies of general aviation (GA) pilot performance are lacking, especially with respect to envisioned future requirements. Two research studies were conducted to evaluate human performance using new technologies. In the first study, ten participants completed the Modified Rhyme Test (MRT) in an effort to compare the intelligibility of two text-to-speech (TTS) engines (DECtalk and AT&T's Natural Voices) as presented in 85 dB(A) aircraft cockpit engine noise. Results indicated significant differences in intelligibility ($p \leq 0.05$) between the two speech synthesizers across the tested speech-to-noise ratios (S/N) (i.e., –5 dB, -8 dB, and –11 dB S/N) with the AT&T engine resulting in superior intelligibility in all of the S/N. The AT&T product was therefore selected as the TTS engine for the second study.

In the second study, 16 visual flight rules (VFR) rated pilots were evaluated for their data link performance using a flight simulator (ELITE i-GATE) equipped with a mixed-modality simulated data link within one of two flight conditions. Data link modalities included textual, synthesized speech, digitized speech, and synthesized speech/textual combination. Flight conditions included VFR (unlimited ceiling,

visibility) or marginal VFR (MVFR) flight conditions (clouds 2800 feet above ground level [AGL], three miles visibility). Evaluation focused on the time required accessing, understanding, and executing data link commands. Additional data were gathered to evaluate workload, situation awareness, and subjective preference.

Results indicated significant differences in pilot performance, mental workload, and situation awareness across the data link modalities and between flight conditions. Textual data link resulted in decreased performance while the other three data link conditions did not differ in performance. Workload evaluation indicated increased workload in the textual data link condition. Situation awareness (SA) measures indicated differences in perceived SA between flight conditions while objective SA measures differed across data link conditions.

Actual or potential applications of this research include guidance in the development of flight performance objectives for future GA systems. Other applications include guidance in the integration of automated voice technologies in the cockpit and/or in similar systems that present elevated levels of background noise during normal communications and auditory display operations.

ACKNOWLEDGEMENTS

I would like to acknowledge several individuals whose expertise, professionalism, and guidance contributed largely to this work. First, I would like to thank Dr. John G. Casali for his guidance and support towards the development and conduct of this research. I would also like to thank Gary S. Robinson for his invaluable insight and direction in the operations of the Auditory Systems Laboratory, and who served as the 'go-to guy with the answers' when I have run into problems or have had questions. Thanks also to Drs. Brian M. Kleiner, Thurmon E. Lockhart, and Antonio A. Trani for their advice, assistance, and expertise and for serving on my advising committee.

I would also like to thank the United Parcel Service for their monetary support in the form of the UPS Fellowship as well as the National Aeronautics and Space Administration (NASA) for funding the flight simulator. Will Vest deserves thanks as well for his electrical talents in the development and integration of various systems and auxiliary hardware in support of my experimental needs for the flight simulator. Special thanks also to Michele Marini, whose knowledge and expertise has benefited me greatly as I waded through the world of statistical analysis.

Lastly, I would like to thank my friends and family. Thanks to my good friend Stacey Lester for his steadfast support and friendship throughout my life. Thanks also to my fellow students and officemates: Jason Saleem, Chuck Perala, Thomas Davis, Bill Penhellagon, and Brian Valimont. Special thanks to my parents, Ralph and Sandra Lancaster, for their continual support and love.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

INTRODUCTION

General Aviation and the National Airspace System

In order to operate safely, general aviation (GA) pilots have long dealt with the need to construct and maintain an accurate mental model of their aircraft and its relation to other aircraft operating in the general vicinity. Such a need is of utmost importance in the safe operation of our National Airspace System (NAS). Thousands of GA aircraft must safely operate in and around locations that range from the remote to the urban, and within traffic that ranges from other GA aircraft to large commercial transports and military aircraft at various densities. To add to the complexity, GA aircraft operate from airports whose capabilities range from a single, non-towered airstrip with little traffic to multiple-runway, thousands of operations a day air-traffic-controlled (ATC) airspaces. As aircraft become more complex in their capabilities, both in performance and in the information that they present to the pilot, there is a need to ensure that pilots can satisfactorily perform their duties, in terms of both safety and compliance to established aviation procedures.

Technological complexity is not limited to the aircraft. Future iterations of the NAS will attempt to harmonize technological innovations of aircraft with the airspace in which they operate. One such iteration is the Small Aircraft Transportation System (SATS). Utilizing the latest innovations in avionics, communication, and automation, SATS attempts to offer enhanced services to users of the NAS that include Higher Volume Operations (HVO)[improved throughput with respect to clearance from runway/airport obstructions], Integrated Fleet Operations [data link and other

communication systems], Lower Landing Minimums (LLM)[landing capabilities below traditional ceiling levels, associated with non-precision approaches] and increased single-pilot safety and mission reliability. In order to ensure that users can operate safely and efficiently in this near-future operational context, it is imperative that research initiatives are explored that address the human element in the system; that is, what are their capabilities, expectations, and limitations and how do they relate to various elements that are envisioned for this system?

The new tasks and procedures that are proposed for systems such as SATS may thus affect human mental workload (WL). It is therefore of utmost importance that, within this new operating paradigm, mental workload does not present stresses to the pilot such that he/she is unable to perform his/her piloting duties in a safe and efficient manner (i.e., the resources available to the human are less than the demands required in the situation). There is also a sizable element of time-sharing or selective attention that occurs in current cockpits—pilots must be aware of any and all functions or states of the aircraft, both inside (e.g., instrument panel) and outside (e.g., control surfaces), as well as activities that are occurring in the airspace around them. This critical aspect of an aviator's job, commonly referred to as situation awareness (SA), can be considered an internalized mental model of the current state of the environment around them. Mental model construction and maintenance is no small feat considering the myriad of changing conditions, indicators, and processes that typically occur in flight. It is only through focused research that attempts to simulate these new operating tenets that measures of mental workload and situation awareness can be gleaned.

The i-GATE Simulator

The research described herein attempted to address a specific subset of a SATS-like operating scheme (that of station keeping and compliance with ATC directives) and how current technologies as well as promising ones can affect pilot performance in these conditions. The primary research tool utilized in the investigation was the i-GATE (Integrated General Aviation Training Environment) simulator, which is a personal computer aviation-training device (PC-ATD) that records flight data of variables related to flight (see Figure 1). The simulator will be discussed in detail later in the apparatus section. Please see Appendix A for additional images of the simulator and experimental setup.

Figure 1. The i-GATE flight simulator.

Current Research

Some future GA efforts will include various non-standard flight conditions and support equipment that purport to ensure safety and efficiency. Examples of these are operations within novel glide slopes and use of controller-pilot data link systems (CPDLS). Limited GA operational investigations have been conducted within non-standard glide slopes (e.g., Lancaster, Saleem, Robinson, Kleiner, and Casali, 2003), but no locatable research has evaluated GA single-pilot operations utilizing data link or variations of data link (e.g., speech and/or textual format of data link information). It is therefore imperative that these proposals, as mentioned, are investigated as to their effects on and usability by human operators, especially the introduction of automation into areas that have traditionally been under the purview of humans (e.g., reduction/limiting of the traditional radio 'party-line'). Literature review, interviews with subject mater experts (SME), and ecological observations have been conducted to help determine what aspects of the task warrant attention, what the appropriate measures to take are, and what elements constitute or contribute to unsatisfactory performance. Current methods in flight performance assessment include measures of workload and situation awareness. There has been some discussion in the literature concerning specific SA measures and their correlations with each other (or lack thereof) as well as with workload; see Endsley and Sollenberger, 2000. One product of the current research is to provide additional data regarding the use of standard techniques in SA measurement (both subjective and objective). The results provide useful information to designers and planners regarding pilot performance in future GA systems.

BACKGROUND

The National Airspace System and Human Factors

The three classes of aviation in the United States (U.S.) are military, commercial (airlines, cargo carriers), and everybody else. 'Everybody else' means general aviation. General aviation includes all varieties of powered aircraft, including helicopters, government, and other non-military aircraft. General aviation pilots consist of the newly licensed pilot who might fly on some weekends in a small airplane to the corporate pilot who flies a sophisticated jet each day and travels all over the world. General aviation, therefore, represents a highly variable segment of aviation with respect to pilot skills, aircraft flown, and equipment utilized.

However, there is an ever-increasing demand for air travel, and, concomitantly, the management of air traffic has become difficult. As a result, constant reports of airline delays and congestion are in the news. The US Department of Transportation (DOT) found that air traffic has grown 12% from fiscal year (FY) 1996 to FY 2000. Additionally, the report states that air traffic operations are expected to increase another 30% by 2011 (DOT, 2000). Today's Air Traffic Management (ATM) system has grown in an evolutionary manner over the past 60 or so years to its present level and, in the opinion of the author, its standard of high safety. In pursuit of ensuring and maintaining current safety capabilities in the increasingly crowded airspace of the NAS, various research programs utilizing extensive machine intelligence have been conducted in support of the human element; that is, Air Traffic Control Specialists (ATCS) as well as the aircraft operators (pilots) in their duties of planning, coordination, communication,

and control. However, any investigations into capacity must also consider safety, for these elements go hand-in-hand. In support of the high safety standards required, it has been related (Blom, 1992) that mathematical collision risk models for controlled air traffic in route networks should be employed, thereby fostering numerical evaluation. Human factors investigations can provide data for these evaluations.

In support of these models, human factors studies of ATC form two main categories. Some studies belong to programs of 'continuous work' that extend over several years and utilize dedicated ATC facilities and/or in-house resources or may even employ contractors under the auspices of national or international agencies. Other studies apply the demonstrated knowledge and expertise of a relevant contractor or academic department to a given ATC problem for an abbreviated time, not becoming involved in the larger picture of ATC issues. This dichotomy has characterized the human factors contributions to ATC in most situations throughout the world (Hopkin, 1995). Human factors and its relation to psychology applies its tenets to specific measures and methodologies within experimentation rather than on the application of psychological theories and/or constructs to the studies themselves or to their interpretation. The international nature of ATC, the universal demands for the safe and efficient handling of more traffic coupled with technological advances, and the quest for effective uses of automation have combined to present human factors problems in ATC and a consequent need to coordinate human factors efforts to avoid duplication (Hopkin, 1995).

Future ATC systems will incorporate new technology, computing, automated assistance, and strategic methods in both GA and ATC; additionally, both will utilize

human operators for the foreseeable future. Within such automated operating regimes, however, human factors investigations will need to account for the changing role of the operators, be they ATC or pilots. Likely changes in mental workload resulting from alternating periods of vigilance or the need to stay alert for long periods of time while few events are occurring (e.g., operations within class E airspace, see Table 1) to periods of high workload resulting from rapid event rates (e.g., operations within class B airspace) will need to be considered and supported within the cockpit and in the control centers. Both pilots and ATC will continue to need training, especially as technologies are introduced, to maintain knowledge and skills. As Hopkin (1995, p. 10) further relates,

> In one sense, the objective of human factors contributions to ATC is the same as that of ATC itself: namely the safe, orderly and expeditious flow of air traffic; a secondary but essential objective of human factors is to ensure that tasks are well-matched with human skills and abilities, thereby fostering not only safe operations, but a satisfying and worthwhile job for controllers.

The Aviation Safety Research Act of 1988 increased awareness of the possibilities of human factors in the NAS. It required the Federal Aviation Administration (FAA) to expend a finite portion of its annual budget on human factors related to systems under development. The *National Plan for Aviation Human Factors* (FAA, 1991) was a product of this law. It was a very comprehensive document that theoretically defined the human factors research required for the present time and the foreseeable future.

TABLE 1

NAS Airspace Designations

	Class A	Class B	Class C	Class D	Class E	Class G
Altitude	> 18K ft.	Variable	Variable	Up to 1200 ft. AGL	Up to 18K ft.	Variable
Airspeed	Unlimited	250 knots max	200 knots max	200 knots max	250 knots max	250 knots max
ATC Service	Controlled	Controlled	Controlled	Part-time control	Controlled	Uncontrolled
VFR Visibility	N/A	3 statute miles (s.m.), clear of clouds	3 s.m., 1K ft. above, 500 ft. below, 2 K ft. from	3 s.m., 1K ft. above, 500 ft. below, 2 K ft. from	Variable	Variable

Since human factors engineers wrote the plan, the FAA has initiated some of the proposed improvements. The FAA then, in 1995 (p. 12), in cooperation with other agencies, developed a revised and consolidated plan, stating the following:

> Human-centered automation research focuses on the role of the operator (active or passive) and the cognitive and behavioral effects of using automation to assist humans in accomplishing their assigned tasks for increased safety and efficiency. The research in this arena addresses the identification and application of knowledge concerning the relative strengths and limitations of humans in an automated environment. It investigates the implications of computer-based technology to the design, evaluation, and certification of controls, displays, and advanced systems.

As ATC continues to evolve in concert with the aircraft it controls, evaluation and certification of advanced systems will be required. The higher the system level at which measurements are taken with respect to the validation of new technologies, the more encompassing the performance criteria must be. Another issue is the degree to which experimental manipulation is required to elicit potential (and perhaps adverse) interaction effects. A comprehensive simulation of an ATC/pilot operational environment can

enable a variety of individual and system-level performance measures to be taken under controlled experimental conditions, and, as Stager (2000) relates, critical performance measures may then be compared against operational data.

Operational testing with the use of real-time simulation affords the ability to control the parameters of critical variables. The representativeness of variables, subjects, and setting (ecological validity) can be viewed as the major component of external validity (i.e., generalizability of findings). These qualities make flight simulation attractive to human factors engineers. The i-GATE simulator, certified by the FAA for instruction and training, has the capability to support such experimentation. Its glass cockpit also fosters desirable external validity, since many new aircraft (e.g., Cirrus, Lancair), and arguably all future SATS-like aircraft, are or will be equipped with such technology.

Human factors investigations of cockpits or ATC systems have relied extensively on simulation, which can be a costly research tool in terms of resources and funding (Hopkin, 1995). Indeed, it is a major commitment to simulate either a cockpit or an ATC system, without trying to simulate both together and the interactions between them, especially in what can only be termed a paradigm shift in operations dictated by SATS-like regimes. Most studies of cockpits or of ATC have each included only those limited aspects of the other that appeared essential to obtain valid findings (Hopkin, 1995). Developments such as data link require more human factors consideration of the communications between air and ground, which must be fully compatible with the equipment and procedures that exist in cockpits and ATC systems and must foster safety

and efficiency (Hopkin, 1995). This consideration is the impetus behind the current research.

One of the primary tenets of human factors is that a relationship exists between the efficiency with which people operate and maintain equipment and the ultimate effectiveness of that equipment's functioning. Equipment characteristics influence how humans operate and maintain that equipment, and, since these characteristics function as user stimuli, it follows that certain arrangements and qualities of them will optimize efficiency. It is therefore to advantage to conduct investigations of proposed future GA systems, including the manner in which these systems are incorporated into the next-generation aircraft that will serve as transporters, as these factors have definite effects on safety and efficiency.

Data Link

The creation of a 'free flight' or SATS-like regime involves the implementation of the controller-to-pilot data link communications system. Data link technology will allow ATC to replace some voice communication with digital transfers of information directly to the flight deck (Latorella, 1998). Data link represents but one major change on the horizon for aviation. Examples of new or future systems that need to be considered in the flight deck data link development process include the following: Traffic Alert and Collision Avoidance System (TCAS), Low Level Wind Shear Alert System (LLWAS), Global Positioning System/Global Navigation Satellite System (GPS/GNSS), and the Electronic Library System (ELS). Further systems that warrant attention are the Automated En Route Air Traffic Control (AERA) Automation System, Automatic Dependent Surveillance (ADS), and the Advanced Automation System (AAS). Data link

will continuously evolve into the NAS as systems become available and economic considerations permit. Data link avionics must be made available for a variety of aircraft classes including all commercial, military, and general aviation aircraft. In the commercial aviation arena, for example, many flight deck configurations are unique and will require specific research and avionics. The possible flight deck configurations include electromechanical (first-generation, e.g., DC-3), 'glass' (second-generation, e.g., B-767), and 'glass/fly-by-wire' (third generation, e.g., B-777) (Rehmann, Reynolds, and Naumeier, 1993). The use of data link for ATC communications offers many benefits over that of conventional voice traffic. These advantages include a reduction in miscommunications associated with voice interactions, a reduction in radio frequency congestion, and the potential for direct entry of data into an aircraft's autoflight (flight guidance and management) systems. Additionally, clearance messages will have more permanence in the cockpit (related to the ephemeral nature of audio communications) with the capability to print and/or review messages after they have been received (Rehmann, 1997). What is notably absent from recent investigations of data link are experiments considering GA operations. Very little research has been located that explored data link implementation outside of commercial operations, especially those focusing on the capabilities and limitations of the *single* pilot.

The current mix of flight decks in the NAS, according to Townsend (1992), is about 60 percent electromechanical and 40 percent first generation glass (although this ratio has likely shifted more towards the latter since that publication). This mix is under constant change as airlines upgrade their fleet and GA aircraft are redesigned to include the more efficient 'glass' aircraft (e.g., Honeywell Epic, Avidyne

Flight Max, Garmin). The glass flight deck presents opportunities for integrated data link that electromechanical aircraft do not. Electromechanical aircraft will require retrofitting of systems to support data link functions (Rehmann et al., 1993). The FAA has recently begun mapping a plan for the building of an ATM system for the domestic NAS. The system will use advanced communications, navigation, and surveillance (CNS) technologies to support future global flight planning, ATC services, and aircraft operations. Data link technology is already being used for digital Automatic Terminal Information Service (ATIS) predeparture clearance (PDC) and oceanic ATC services (Rehmann, 1997).

In 1993, the FAA began human factors research efforts into data link. Table 2 outlines the research questions identified. It should be noted that investigations into data link display surfaces, types, and location ranks second only to protocols with respect to importance, and shared data link displays is third. Formats and contents ranks within the top ten, and synthetic voice investigations ranks thirtieth. SA measures are also listed. Clearly, at least when considering these FAA human factors data link concerns, the rationale for the research described herein is supported.

With these data link systems, the pilot typically receives an auditory and/or visual signal of an incoming text-based message from ATC (Harvey, Reynolds, Pacley, Koubeck, and Rehmann, 2002).

The Harvey et al. research showed that the frequency of ATC communications was found to be significantly less than non-data link-equipped crews.

TABLE 2 Data link human factors issues by groups as identified by the FAA. The numbers correspond to the overall importance ranking; **bold** indicates areas wherein the current research might provide useful data (from Tech. Report DOT/FAA/CT-TN93/5, p. 31).

Procedures	
1	**Data Link protocols**
8	Effects of delayed unables
14	Modifications to clearances
15	Function allocation
17	Data distribution
18	**Data link implementation evolution**
24	Emergencies and transitions between emergency and non-emergency states
27	Pilot flying/Pilot not flying procedures
28	**Mixed environment**
36	Negotiations
37	Currently nonexecutable clearances
Errors	
9	Effects of controller errors
11	Opportunities for error checking
12	Pilot detection of other controller errors
16	Communication sequence errors
19	Pilot detection of other pilot errors
23	Error recovery procedure
26	Controller detection of flight crew errors
35	Levels of involvement
42	Proficiency loss
45	Pilot detection of other aircraft errors
Human Interface Design	
2	**Display surfaces, types and locations**
3	Shared displays
6	**Crew alerting mechanisms**
7	Expiration times
10	**Formats and contents**
13	Priority displays
21	Standardization
25	Message displacement
29	Menu design
30	**Synthetic voice displays**
31	Clearance evaluations
32	Recovery from accept/reject errors
33	Definition of inhibit Logic
38	**Link status displays**
39	**Display ordering and response facilitation**
40	Discrepancies
41	Too quiet flight deck
44	Selection of information sources
Situation Awareness	
4	Effects of response delays on controller SA
5	Crew information transfer
20	**Data link integration with other cockpit technologies**
22	**Party line compensation**
34	Situation awareness recovery
43	**Independent confirmation**

Additionally, and perhaps more importantly, while data ink use decreased ATC-to-pilot vocal communications, the cognitive demands placed on pilot *crews* utilizing data link was increased, requiring them to interact more in an effort to understand what the messages meant, and how that affected their strategies, intentions, and actions (Harvey et al., 2002). One wonders how the *single* pilot would respond to data link implementation, as a focus of SATS-like operations is on the capability of the *single pilot* to operate effectively within the NAS, and this question serves as a further impetus for the current research.

The transmission and understanding of information have been quite extensively studied in research on communications in aviation. Topics have included errors, presentation of information, tasks, language and vocabulary using auditory or visual data, and appropriate levels of detail (Hopkin, 1995). Because of the volume of radio traffic and its transmission quality in some flight environments, and the heavy activity within the flight deck during critical periods of flight (e.g., approach), the radio can sometimes be a poor form of communication. Speech between ATC and the flight deck fulfills many functions. The judgments and assessments that pilots and controllers make about each other, related to such aspects as their professionalism, ability, confidence, and evident familiarity with tasks and messages, are "based largely on the content, pace, phraseology, consistency, standardization, courtesy, and felicity of expression of the spoken messages between them" (Hopkin 1995, p. 27). Hopkin further notes that pilots make judgments about the competence and reliability of the ATC service they are receiving, and request clarification, confirmation or supporting evidence accordingly. Similarly, ATCS make judgments about the pilots with which they communicate; they

may check frequently that their instructions are being obeyed or require more transition states be reported if they believe a pilot is inexperienced or unfamiliar with local procedures. The judgments may sometimes be unproven, but speech between pilot and controller conveys much more than the quantitative content of the spoken messages. If this avenue of communication is curtailed through usage of data link, so is the basis for such judgments (Hopkin, 1995).

In studies involving airborne data link, each uplink (except initial contact) has typically required a 'WILCO' or 'UNABLE' response in order to complete the transaction (Rehmann, 1996, 1997). The total amount of time required to access and respond to data link messages is important from the perspective of the ATCS. Controllers are accustomed to rapid radio response from pilots and are reluctant to use data link when there are long delays to receipt of WILCOs (Rehmann, 1996). As the Rehmann research has focused solely on commercial operations utilizing flight crews, there is a definite need to evaluate data link usage, including response times, within the GA domain, where there typically are not flight crews, but a single pilot.

Even though the pilot is legally responsible for the safety of his/her aircraft and its passengers, and the ATC is legally responsible for the safety of the air traffic control instructions he/she provides, Hopkin (1995) notes that such boundaries may blur in the presence of data link capabilities. When both the pilot and the controller are implementing air traffic control instructions that are presented on screens or through another modality in the cockpit and in the air traffic control workspace, but have been derived from software in the air or on the ground, the issues of legal responsibility become quite complex. Hopkin (1995) further relates that the "ultimate reason for the

retention of humans in aircraft cockpits and in the air traffic control systems may be their legal responsibilities rather than considerations of ATC, of human factors, of technology, or of aviation"(p. 28).

Ideally, the results of human factors studies and experiments are (or should be) integrated into analytical models that seek to simulate system operation and performance. There are fears, however, that the increased automation envisioned for future GA systems (such as data link systems) will be met with some resistance. Within the recent history of automation integration, especially when coupled with known performance effects related to automation (i.e., complacency and monitoring considerations), one can expect integration issues to arise in this regard. Automation can directly impact situation awareness, for example, through several mechanisms: (1) changes in vigilance and the complacency associated with monitoring, (2) assumption of an increasingly passive role instead of an active role in system control (e.g., autopilot maintenance in 'highway in the sky' [HITS]-equipped aircraft), and (3) changes in the quality and form of feedback provided to the human operator (Svennson, 1997). (The concept of situation awareness and its relation to the current research will be discussed in some detail later.) Automation concerns will soon be brought to the fore as these systems are tested, and represent a factor to consider in the safety and efficiency of a decentralized ATC/ATM system. Although many issues remain unresolved, global positioning sensors and miniature inertial and rate sensing instruments combined with conventional air data systems will soon support inexpensive integrated measurement systems for GA aircraft that will provide accurate measures of such indices as linear and angular positions and velocities as well as airspeed, angle of attack, and side slip (Thompson, 2000). Along with

technological advancements and the concerns they may bring, one also needs to consider how these advancements may aid pilots of future aircraft in their piloting tasks. Graphical weather systems may reduce voice communication (and thus error potentials) as well as cockpit workload. Further, and as discussed, data link can increase the safety and efficiency of a decentralized ATC/ATM structure through reductions in communication errors and through providing for increased data flow between aircraft and ground facilities. This is quite a desirable quality for, as Prinzo (1996) relates, in 1993 there were 255 near midair collisions, 38 of which (15%) were the direct result of communication discrepancies. As such, data link may be particularly effective in high-density terminal areas during peak travel times via a more efficient handling of ATC clearances. Preliminary investigations of data link have shown (Phillips 1992, Rehmann 1993, Rehmann, Reynolds, and Naumeier 1993) that the system decreased the number of transmissions (thus promoting more of the 'aviate' in the 'aviate/communicate/navigate' paradigm), reduced demands on pilots' short term memory, and allowed air crews more time to perform critical cockpit tasks while receiving ATC instructions. Phillips also relates that pilots' head-down time was significantly increased in a text-only data link condition, thus reducing 'out-the-window' vigilance, further supporting investigations such as the experiments in this document. Explorations into the effect of advanced flight management systems (FMS) may further support safety and efficiency by fostering quick entry of such indices as airspeed, heading, and altitude commands.

Blom, Stroeve, Daams, and Nijhuis (2001) discuss the development of a stochastic analysis-based methodology that takes an integral approach towards accident risk assessment for air traffic. They state that views of human reliability have shifted

from a ‘context-free error centered approach’, in which failures of human information processing are used for unreliability modeling, towards a ‘contextual perspective’ in which human internal state, strategies, and the environment affect human actions in a contextual perspective. In this regard, the modeling of safety critical human actions is suggested in relation to the other activities engaged by the operator and the environment. For a proper description of human reliability, it is necessary to include the cognitive processes that underlie the actions of humans. This leads to a comprehensive model of operator performance (Blom et al., 2001). Hence, the output of human factors investigations can and should be utilized as input into human reliability models.

Speech Technology

Text-to-speech (TTS) systems. As data link capabilities have often been suggested for incorporation into next-generation GA aircraft, such systems of course need to be evaluated for their impacts on human operators. Many studies (Begault 1993, 1998; Begault and Wenzel, 1992; Begault and Pittman, 1996) have demonstrated the desirable ability of three-dimensional (3-D) auditory displays to foster situation awareness with respect to ground- and airborne-based traffic. Further, the capabilities of TTS systems have experienced sizable improvements within recent years in their production of realistic-sounding (i.e., human sounding) speech; however, no locatable research has surfaced in which these newer systems are evaluated. Therefore, it seems to advantage to investigate the latest speech technology for possible use in aviation communication systems.

A practical need has always existed for devices that have the capability to produce as well as understand spoken language automatically without human intervention. The technology has evolved to produce specialized microprocessor-based speech processing devices that can easily be integrated into numerous computer-based systems in the support of user-machine communication. Other integrations, however, may prove difficult.

Speech is the most natural means of communication between humans. It is automatic, requires little conscious effort or attention, and creates few, if any, demands while other tasks are carried out concurrently, especially tasks which require active use of the hands or eyes in the demanding environs and situations that are present in the cockpit. A potential use of speech is as an interface with computers, systems that have traditionally been interacted with via keyboards, mice, and/or screens (Greene, Logan, and Pisoni, 1986). One must understand, however, the relative merits of each type of system, what constitutes a successful system, and the implications of poor design with respect to auditory displays that utilize speech synthesis and/or digitized speech.

Early attempts at text-to-speech synthesis assembled clauses by concatenating (i.e., connecting through a link or series) recorded words. This technique produces extremely unnatural-sounding speech. In continuous speech, word durations are often shortened and coarticulation effects (i.e., union of words) can occur between adjacent words. There is also no way to adjust the intonation of recorded words, which is an important element to the understanding of words. A huge word database is required, and words that are not in the database cannot be pronounced. The resulting speech sounds choppy.

With a text-to-speech system, virtually any computer can generate spoken output from a string of characters and can therefore provide users with novel speech displays instead of the more traditional screen. In some applications, the display may significantly reduce users' workload and increase operators' efficiency in information retrieval from a computer. In other applications, such as airborne data link, TTS systems may provide entirely new methods for data retrieval as well as data manipulation, and these methods need to be investigated.

Some TTS systems (e.g., DECtalk) produce voice output using various synthesis-by-rule techniques: techniques that generate speech through attention to a series of rules, which are used to create utterances on demand. Typically, these systems are highly sophisticated (and thus expensive) and consist of a number of modular subsystems, each of which has a special set of rules. Initial typed input is first converted into ASCII code, and, in most current systems, the code is further processed through several modules, which serve to produce a detailed phonetic description. In many of these systems, the analytic process involves the determination of the underlying phonemic, syllabic, morphemic, and syntactic form of the input message, as well as adjustment of the input when numerals, abbreviations, and special symbols are present (Greene et al., 1986). After basic module operations are complete, any word that has not been analyzed is processed through a set of letter-to-phoneme rules. Once the text has been converted into a phonetic transcription, other modules, typically containing detailed phonological, pitch, stress, and timing adjustments, operate on this representation. Additional rules are included in an effort to make the speech sound 'less mechanical.' Other modules focus

on rules that serve to disambiguate similar-sounding words such as ‘read’ (which can be pronounced like ‘red’ or ‘reed’); see Greene et al., 1986.

After these analyses, the inputted text is converted into spoken output, a process that is also modular. Several modules are used to specify the manner in which each speech sound is to be pronounced, how certain other speech sounds are to be modified by specific contexts, and where stress is to be placed. Quite obviously, the more detailed the rule system, the more nearly the synthesized results mirrors natural speech. All of the parametric information accumulated at that point is then input to a digital speech synthesizer and a speech waveform is generated. Finally, the samples are converted to analog (again, via a digital-to-analog converter) and are low-pass filtered (Greene et al., 1986).

In systems that utilize synthesis-by-rule techniques, technological advances are particularly notable in their speech output. Such systems require code conversions through several modules in which various mathematical analyses are conducted to result in detailed phonetic descriptions. These ‘detailed rule sets’, as described by Greene, Logan, and Pisoni (1986), require increased processing power; the more detailed the rule set, the more natural the speech. It follows that today’s increased computing power can result in improved, more natural-sounding speech output. Because the processing speed and memory capacities that are available today were not available in the 1980s and 1990s, the capabilities of earlier speech synthesis systems resulted in speech output that was severely ‘impoverished’ (i.e., choppy, lacking in prosody and acoustic cues).

Other advances in TTS technology include concatenative systems (e.g., AT&T Natural Voices). These engines concatenate parameterized units of natural speech to

produce speech output, typically using linear predictive coding (LPC) as the synthesis method (Venkatagiri, 2003). As natural speech is sampled and concatenated, speech output from these systems can be expected to be very natural, at least when compared to a system that utilizes synthesis-by-rule (e.g., DECtalk). Diphones exist as the natural speech unit in these systems, and 'naturalness' is retained by diphone extension between the centers of adjacent phonemes. Other concatenative techniques utilize waveform-based methods. Such systems enjoy greater power in controlling various prosodic variables over LPC synthesis (Venkatagiri, 2003).

Digitized speech systems. When natural speech is recorded onto audiotape (usually digital audio tape [DAT]) using a microphone, tape recorder, and a computer using an analog-to-digital converter, it then becomes digitized speech (also known as 'stored speech'). The actual process involves sampling the speech waveform at a rapid rate and storing the samples in digital form. Typically, from 8,000 to 10,000 samples are taken for every one second of speech (Greene et al., 1986). The samples are then stored in computer memory as a series of numerical parameters. Thus, a five second sentence will have at least 40,000 samples associated with it; each of these samples will be stored digitally in the computer. As such, for long passages, the storage needs are enormous. Thankfully, today's memory prices are cheap enough that the storage need is not as much of an issue as it was only a few years ago.

There is a good reason to use stored speech. All the digital samples can be retrieved from the computer memory and then reconverted to analog form using a digital-to-analog converter. The process reproduces the speech that was originally recorded with little or no degradation or effects on intelligibility. Although there may be some loss in

speech quality due to the sampling rate and the number of bits used to code the speech waveform, the resulting speech quality is acceptable and often sounds better than speech transmitted over the telephone (Greene et al., 1986). However, if the message needs to be changed or updated, the entire process must be repeated. Thus, stored speech is useful for very limited message sets, such as the letters of the alphabet, the digits 0 to 9, or a very small vocabulary of key words or instructions. One wonders whether ATC commands fulfill these conditions, since there are a relatively small number of commands and directives that are utilized, especially when compared with normal conversation. When the vocabulary becomes very large and the potential set of messages is theoretically unrestricted, a voice output system using stored speech becomes impractical and prohibitively expensive. Further, when individual stored items are combined into word strings without any additional processing or smoothing, the speech that results lacks normal pitch and intonation (prosody); listeners often describe speech in this condition as unnatural and 'mechanical sounding.' It is not surprising, therefore, that the intelligibility of this kind of connected speech is often quite poor, even though the intelligibility of individual words is typically quite high (Greene et al., 1986).

With respect to ATC commands, Cotton and McCauley found that when a system simulates human communications (related as a Navy ATC training system), a natural-sounding voice, using digitized human speech, was preferred (1983). However, no research investigations could be located with respect to preference within the context of an operational GA cockpit, and certainly none could be found comparing mature synthesized systems with digitized ones.

Cognitive costs of natural speech vs. synthesized speech. Paris, Thomas, Gilson, and Kincaid (2000) caution there are cognitive processing costs associated with speech comprehension of TTS-synthesized speech, relative to the comprehension of natural speech. They note the results of research investigations in this arena; that speech synthesizer output places increased burdens on perceptual and cognitive resources during the process of comprehension, and that performance decrements have been discovered at many stages of processing, from phonemes to paragraphs (Paris et al., 2000). This burden, termed 'cognitive cost,' is paid through the currency of performance decrements in many stages of processing (e.g., from phonemes to paragraphs) resulting in undesirable conditions such as increased workload and decreased performance in human-in-the-loop systems that attempt to incorporate such TTS engines.

Several factors are posited to account for processing speed differences between synthetic and natural speech. First, differences exist in the amount of information conveyed by natural and synthetic speech at the phoneme level. Synthetically generated phonemes are 'impoverished' relative to natural speech because many acoustic cues are either poorly represented or not represented at all (Paris et al., 2000). Another important difference is in the extent to which prosodics are appropriately modeled. Cues inherent to prosodics provide perceptual segmentation and redundancy, speeding the real-time processing of continuous speech; they guide expectancies, cause search processes to end when contact is made between an acoustic representation and a cognitive one, and influence the actual allocation of processing capacity in terms of power, temporal location, and duration (Paris et al., 2000). Synthetic speech systems, however, are limited in their prosodic capabilities, particularly with respect to the emulation of

appropriate stress and intonation patterns. Correct usage of 'contrastive stress' (unspecified) requires an appreciation of the meaning of a particular utterance based on "accurate parsing of its syntactic and semantic components" (Paris et al., 2000, p. 422). Prosody in TTS synthesizers is said to be generally limited to the addition of pitch contours to phrase units marked by punctuation. Because these variations are implemented by rule sets, the resulting prosodic markers are less robust than for human speech and may even be incorrect (Paris et al., 2000). Again, however, newer TTS systems may or may not maintain these undesirable qualities.

Lexical complexity may also affect intelligibility. If the acoustic cues of a word are not intelligible or are misleading with respect to the prosodic or segmental patterns they indicate, then listeners may not be able to make use of their knowledge of morphological structure to improve recognition (Francis and Nusbaum, 1999). As a result, those synthesizers that are accurate at reproducing acoustic cues of natural speech may be said to further facilitate recognition by fostering listener use of his/her full range of pattern knowledge of spoken language. This contention is supported in their research, wherein listeners were able to recognize polymorphemic words more accurately than monomorphemic ones; even though the former can be said to be less familiar and less common, listeners could clearly use the structural constraints provided by morphological structure to aid in word recognition (Francis and Nusbaum, 1999). The researchers note that, for low-quality synthesizers such as VOTRAX (an antiquated system dating from the 1970s), recognition is difficult due to the aforementioned constraints; conversely, higher quality synthesizers, such as DECtalk, perform relatively well at producing two-syllable words, and morphological complexity does little to provide assistance in

recognition. These results are related to be a function of the accurate natural speech production enjoyed by users of such systems (Francis and Nusbaum, 1999).

***Speech synthesizers*.** Presented in this section is a brief list of speech synthesis systems that exist on the market as well as their capabilities. The type of synthesis varies from one TTS engine to another and can be one of three types*: formant-based*, *articulation-based*, or *concatenative.* Several of these systems have been investigated in the literature as relates to intelligibility; these research endeavors as well as their results will be discussed in another section.

- *MITalk*: The MITalk system was initially designed as a research tool. It was implemented on a DECSYSTEM-20 computer at the Massachusetts Institute of Technology (MIT, hence the name), and was the product of a 10-year effort to convert unrestricted English text input into high-quality speech output (Greene et al., 1986). MITalk consists of a number of program modules which first analyzed the text input in terms of morphological composition and performed a lexical look-up operation to determine whether or not each morpheme (i.e., a meaningful linguistic unit consisting of a word, such as *man*, or a word element, such as *-ed* in *walked,* that cannot be divided into smaller meaningful parts) was present in a 12,000-item dictionary. If the morphemes composing the words were not found in the dictionary, another module containing approximately 400 letter-to-sound rules was used to arrive at a pronunciation of the text. In addition, sentence-level syntactic analysis was also carried out in order to determine prosodic (i.e., of or relating to the metrical structure of verse)

information such as timing, duration, and stress. The parameters resulting from these analyses of the text were then used to control a formant synthesizer. The MITalk system runs in about 10 times real time due to the time required for I/O operations (Greene et al., 1986).

- *Prose 2000*: From Telesensory Systems, Inc., the Prose 2000 and other Prose products are available from Speech Plus, Inc. The first prototype was based in part on the MITalk-77 system, but only used a 1,100-unit dictionary for lexical look-up; it omitted the parsing system, and replaced the MITalk fundamental frequency module with a 'hat and declination' procedure (unspecified).
- *DECtalk*: DECtalk is a stand-alone TTS system produced commercially by Digital Equipment Corporation (DEC) and recently (2001) sold to Force Corporation. The device was designed to produce high-quality synthetic speech, and is considered by many the best offering in this respect. It also has a wide range of useful features, such as the diversity of available voices, the flexibility of a user-defined dictionary, and standard telephone interfaces. The DECtalk Software development kit consists of a shared library (a dynamic link library on Windows NT), a link library, a header file that defines the symbols and functions used by DECtalk Software, sample applications, and sample source code that demonstrates the API. The DECtalk software supports nine preprogrammed voices: four male, four female, and one child's voice. Both the API and in-line text commands can control the voice, the speaking rate, and the audio

volume. The volume command supports stereo by providing independent control of the left and right channels. Other in-line commands play wave audio files, generate single tones, or generate dual-tone multiple-frequency (DTMF) signals for telephony applications. DECtalk technology exists within many commercially available speech synthesizers.

- *Street Electronic Echo*: The Echo TTS system is an inexpensive system manufactured by Street Electronics and is designed primarily for the computer hobbyist market. Using an algorithm developed at the Naval Research Laboratory, text is converted into allophonic (i.e., a predictable phonetic variant of a phoneme) control codes which are then converted to speech using linear predictive coupling (LPC) synthesis by use of a Texas Instruments TMS-5200 chip.
- *Votrax Type'n'Talk*: the Votrax system is a relatively inexpensive TTS product manufactured by Votrax Inc. Text is converted to phoneme control codes by a text-to-speech translator module. These codes serve as input to the SC01 phoneme synthesizer chip, which utilizes formant synthesis techniques to produce speech. All speech is generated by rule.
- *Berkeley Systems Works*: The Berkeley system is a prototype device that used the General Instruments SP1000 chip to carry out LPC synthesis of allophonic segments generated by a set of proprietary rules.
- *Infovox SA 101*: The Infovox SA 101 TTS system is another stand-alone unit based on synthesis rules developed for Swedish and English by Carlson and Granstrom. It was developed in Sweden at the Royal Institute

of Technology an was commercially implemented by Infovox AB. The most distinctive feature of this system is its multilingual capability.

- *Lucent Technologies/Bell Labs TTS System*: The Bell Labs Text-to-Speech system (TTS) has various applications including reading electronic mail messages, generating spoken prompts in voice response systems, and as an interface to an order-verification system for salespeople in the field. TTS is implemented entirely in software and only standard audio capability is required. At present, it contains several components, each of which handles a different task. For example, the text analysis capabilities of the system detect the ends of sentences, perform some rudimentary syntactic analysis, expand digit sequences into words, and disambiguate and expand abbreviations into normally spelled words, which can then be analyzed by the dictionary-based pronunciation module (“Bell Laboratories,” 2002).
- *IPOX All-Prosodic Speech Synthesizer*: The main data structure in IPOX is a metrical tree, the nodes of which are complex feature structures. This metrical tree is assigned by parsing input text using declarative constraint-based grammars. Each node in the metrical representation is then assigned a temporal domain within which its phonetic exponents are evaluated. Within the syllable, heads are evaluated before non-heads, allowing metrically weak constituents (e.g. onset, coda) to adapt to their strong sisters (rime, nucleus), with which they overlap. Across syllables, the order of interpretation is left-to-right, so that each syllable is "glued" to the previous one. After all phonetics exponents have been evaluated, a

parameter file for the Klatt synthesizer is generated. The current version of IPOX runs under Windows on a 486PC equipped with a standard 16-bit sound card. Graphics are used to display analysis trees, phonetics parameters as well as audio output waveforms ("Speech Synthesis," 2002).

- *Telcordia Technologies' Hybrid ORATOR II*: The Hybrid ORATOR® II Speech Synthesizer from Telcordia provides the tools for high quality, highly accurate telephone access to database-driven information services through advanced text-to-speech synthesis. The Hybrid ORATOR II synthesizer achieves near-human speech quality. Telcordia's Listings Preprocessing software converts telephone company listings into "natural language order," with exceptional accuracy, often reducing the error rates associated with unidentified acronyms, idiosyncratic abbreviations, and incorrect word ordering, by a factor of 20 or more. The software converts and corrects listings from the formats of all major listing vendors.
- *AT&T's Natural Voices*: AT&T Natural Voices' TTS Engine can uniquely support the addition of many languages to any and all applications, including U.S. English, German, Latin American Spanish, U.K. English, Castilian Spanish, Brazilian Portuguese, French, and Canadian French. All editions of the TTS engine include both a female and male U.S. English voice and support SAPI 4.0, 5.0 and 5.1, the SSML component of VoiceXML, and JSAPI interface standards. The Server and Desktop editions of the AT&T Natural Voices' TTS Engine support the creation of

unique customized voices for businesses interested in extending corporate image or brand via the TTS output of their enterprise or customer-facing applications. AT&T's Natural Voices is a concatenative TTS.

See Table 3 for a brief listing of each synthesizer's strengths and weaknesses.

TABLE 3

Speech Synthesizers: Strengths and Weaknesses

Speech Synthesizer	Strengths	Weaknesses
MITalk	Extensive R&D	Outdated
Prose 2000	Built on MITalk prototype	Small dictionary
DECtalk	Extensive R&D & usage	Price
Street Electronic Echo	Inexpensive	'hobbyist' status
Votrax Type 'n' Talk	Inexpensive	Outdated
Berkeley Systems Works	Extensive R&D	Proprietary
Infovox SA 101	Extensive R&D	Outdated
Lucent/Bell TTS System	Usability of interface	Software implementation
IPOX All-Prosodic System	Graphical depictions	Outdated
Telcordia ORATOR II	Telephonic access	Limited usage
AT&T Natural Voices	Wide application support	No locatable research

Digitized speech equipment. Presented in this section are the results of searches for products/equipment whose purpose is to present digitized (human voice) speech. The list is rather limited, arguably due to the fact that most any playback device that is capable of handling digital media (DAT or otherwise) is capable of presenting digitized speech. Indeed, a few of the systems located serve as both recording and playback devices.

- *TALXWare Digitized Speech System*: TALXWare accepts professionally recorded voice files from analog or digital audiotape as input to the digitizing process. TALXWare can also accept audio files in standard formats such as .wav. Silence Processing capabilities ensures that any

noise present during periods of silence (i.e., between words and phrases) is eliminated, resulting in a crisp, clean script when the various words and phrases are concatenated by the application. TALXWare utilizes proprietary 'ValueVoice' software voicing algorithms for parsing data and properly concatenating the phrases necessary to speak complex formats such as digits (123, one-two-three), values (123, one hundred twenty-three), amounts ($123.00, one hundred twenty-three dollars), dates (012345, January twenty-third nineteen forty-five) and ordinal numbers (123, one hundred twenty-third). TALXWare also allows the application to specify the voice inflection to be used during playback. In this way, a variable spoken at the beginning or in the middle of a sentence can use a flat inflection to preserve the natural flow of the phrase. A variable spoken at the end of a sentence can use a down inflection to convey the end of the statement just as we do in normal conversations. Additionally, virtually any language can be used in a TALXWare application and TALXWare includes ValueVoice algorithms for US English, UK English, French, Italian, Portuguese and Spanish. Finally, the portability of TALXWare allows the same digitized speech to be used on multiple hardware platforms (“Talx,” 2002).

- *Zygo MACAW Series*: The MACAW series can access more than 19 minutes of recording time and have its vocabulary saved on computer disks. It is equipped with a built-in hard drive that can store over 13 hours of recording time. The system has over 40 different personalities; each

personality is quickly attainable and contains the vocabulary and all operation parameters like key pattern, scan type, user accessible functions, etc.

- *Adaptive Communication's ALLTALK*: Alltalk is a portable, battery-powered speech output communication device. Selection of voice output is made with an adaptable, touch sensitive membrane overlay. It supposedly generates human voice quality output; the voice of the programmer is stored in re-programmable microchips. Standard memory capacity is 600 words, which can be expanded to 1200 words with an available adapter (Alltalk 4). The expanded memory (Alltalk 4) permits the user to sequence pictures and store different vocabularies on four levels. Additional vocabularies may be stored using a tape-recorder.
- *DigiVox 2000*: The DigiVox 2000 can be purchased with up to 142 minutes of recording time. The system has the ability to save an unlimited number of voice messages by copying them to a floppy disk using a DigiVox 2000 Disk Drive. Thus, a library of special messages can be built for the user to accommodate different situations.
- *IntroTalker*: The IntroTalker is a lightweight communication device designed to be used as an evaluation tool or a communication aid for those with limited needs. It is easily programmed by speaking into the built-in microphone. The standard module holds two minutes of speech. Additional memory modules can be added to increase this to eight minutes. The standard IntroTalker has 32 keys on 38 mm (1.5 inch)

centres requiring 4 ounces of force for activation. An eight-location operating kit is also available. The system is an oblong box with eight columns of four squares. The idea is to put a picture, symbol or word on a square with an associated message behind it. For example, when the user presses the square which has a picture of a cat on it the word 'cat' is spoken. Scanning IntroTalkers with switch access are also available.

See Table 4 for a brief listing of each speech digitizer's strengths and weaknesses.

TABLE 4

Speech Digitizers: Strengths and Weaknesses

Speech Digitizer	Strengths	Weaknesses
TALXware	Inflection, portability	Cost
Zygo MACAW Series	Many personalities	Limited recording time
Adaptive Comm. ALLTALK	Portable	Extra memory needed
DigiVox 2000	Libraries easily formulated	Requires saving on disks
IntroTalker	Lightweight, programmable	Limited recording time

Current flight operations and automated speech. Pilots operating within the airspace of busy airports (i.e., class B or class C airspaces [i.e., larger- and smaller-sized 'busy' airports; for example, Chicago O'Hare and Norfolk, VA, respectively]) are accustomed to hearing digitized and/or synthesized speech through the ATIS, and as such can be said to possess some experience with artificial voice intelligibility. Developed in an effort to improve controller effectiveness and to reduce frequency congestion, ATIS is available in selected high frequency terminal areas. ATIS is prerecorded (digitized) or is synthesized and is broadcast continually on its own frequency. At larger airports, there may be a single ATIS frequency for departing aircraft and another for arriving aircraft.

ATIS broadcasts are labeled with successive letters from the phonetic alphabet, such as 'Information Bravo' or 'Information. Charlie.' The next letter identifies each new ATIS broadcast. ATIS is updated when airports conditions change or when any official weather is received. Pilots typically write down ATIS information (Willits, 2002).

***Measures of speech intelligibility*.** Many studies whose aim is the evaluation of the intelligibility of speech systems reveal that many differences exist between synthetic and natural speech. The latter tends to be rather poor from a phonetic point of view and the former appears very redundant. This may be due to the aforementioned rule-based synthesis with which synthetic speech is generated, a technique that manipulates only a limited number of acoustic cues of the phonetic representation of the message. Problems may surface due to the decidedly 'mechanical' quality of synthetic signals with respect to individual word recognition and phrase and sentence interpretation.

High intelligibility has, in many cases in literature searches of the subject area, been achieved at the expense of naturalness. Generally, human conversational speech tends to be 'articulatorily imprecise', and consonantal cues tend to be acoustically fuzzy (Delgou, Conte, and Sementina, 1998). The identification of words is accomplished based on syntactical and semantical contextual cues as well as acoustic ones. Sentence and text comprehension is reliant on listener characteristics (e.g., linguistic abilities involved in segmenting and analyzing speech into appropriate units; content-related knowledge; motivation) as well as external factors such as text properties (e.g., length, complexity) and acoustic properties (e.g., speed, pitch); see Delgou et al., 1998. As the intelligibility of rule-based synthetic speech improves and the number of applications for synthetic speech increases, it is likely the naturalness of synthetic speech will become an

increasingly important factor in usage determination; this is imperative within aviation operations. Delgou et al. (1998) relate that all of the currently available metrics for evaluation of acceptability (including naturalness) of systems do not sufficiently distinguish between *acceptability* and simply measuring the ability of listeners to *extract intelligible information* from the signal. Indeed, almost all of the evaluation methods with respect to intelligibility are derived from standardized tests developed many decades ago for the assessment of signal transmission fidelity (mostly during World War II) or for testing comprehension in the hearing-impaired. Extending from these early endeavors, the Modified Rhyme Test (MRT; House, Williams, Hecker, and Kryter, 1965) was developed, along with the Diagnostic Rhyme Test (DRT). The MRT asks listeners to identify the word they heard from among a set of words differing by only one phonetic unit. Both represent the most frequently used methods in the assessment of the intelligibility of TTS systems (Delgou et al., 1998).

This is not to say that other intelligibility measures do not exist. Benoit, Grice, and Hazan (1996) introduced the 'semantically unpredictable sentence' (SUS) test to measure intelligibility at the sentence level. The sentences can be automatically generated using five basic syntactic structures and a number of lexicons that contain the most frequently occurring mini-syllabic words in each language. Thus, the sentence material has an advantage over the MRT and DRT approaches (which focus on phonemes at the initial and final positions) in that it is not fixed, as words can be extracted from the lexicons in random fashion to form new sentence sets each time the test is run. The researchers relate that various TTS systems in a number of languages have been evaluated using the test, suggesting that it is effective and allows for reliable comparisons

across synthesizers provided the guidelines are followed carefully regarding the definition of the test material and actual running of the test (Benoit et al., 1996).

SPIN (Speech In Noise) sentences have been used to control for the effect of context. Originally developed as audiometric speech material, SPIN sentences were designed so that the effects of semantic information at the sentence level were controlled. Control is accomplished through presentation of words in either high- (HP) or low-probability (LP) contexts (e.g., HP: "We're lost, so let's look at the map" versus LP: "I should have considered the map"). Differences between scores for HP and LP words provide an indication of the amount of information provided by the sentence context. The main disadvantages, again related by Benoit et al., is that it is lengthy to administer, only tests a single word category (the last noun in the sentence), and consists of only ten fixed lists and therefore does not provide enough material for large-scale comparative tests, as sentences cannot be used more than once because of learning effects (1996).

Another intelligibility measurement tool is the Hearing in Noise Test (HINT). Nilsson, Soli, and Sullivan introduced an alternative technique to *percent intelligibility* in the form of the *speech reception threshold (SRT)*, which is argued to hold an advantage over the former because the latter is "not subject to floor and ceiling effects" (1994, p. 1085). The SRT is defined as "the presentation level necessary for a listener to recognize the speech materials correctly a specified percent of the time, usually 50%" (Nilsson et al. 1994, p. 1085). Designed primarily as a research tool with which to directly assess the impacts of hearing impairment on communication, this rather interesting approach focuses on varying the level of subsequent stimuli based on a correct or an incorrect response. That is, when an incorrect answer is given, the level of the next stimulus is

increased; the converse when it is correct. The stimuli include 250 sentences cast into 25 phonemically matched and balanced lists. The researchers relate that in this way, the presentation level will approach the listener's individual SRT. Using this procedure, according to the researchers, as few as ten sentences per list will provide measurements that are sensitive enough to detect threshold differences of 2.41 dB in noise (Nilsson et al., 1994). While attractive from many standpoints, this tool is not yet widely used or accepted as the MRT.

Many speech intelligibility investigations have demonstrated that listeners do not always agree on measures of voice quality when using traditional rating scales. Gerratt and Kreiman (2001) related that these findings might be due to an inherent inability of listeners to agree in their perception of such complex auditory stimuli, but they think another explanation lays in the measurement methods themselves—the rating scale judgments. As such, they developed an alternative method in quality assessment called 'listener-mediated analysis-synthesis,' which appears to be more externally valid than other measures, and thus has implications for cockpit auditory displays. In this approach, listeners explicitly compare synthetic and natural voice samples, but take advantage of synthesizer features in the comparison: they adjust the parameters of the synthesizer to create auditory matches to voice stimuli. The researchers suggest this method replaces the traditionally unstable internal standards for qualities such as 'breathiness' and 'roughness' with externally presented stimuli (Gerratt and Kreiman, 2001). The analysis-synthesis task is said to provide the same theoretical advantages as the traditional anchored protocol, in that listeners explicitly match reference and test stimuli. However, the analysis-synthesis technique provides a much finer scale resolution, allowing listeners

to create a very close match to the perceived quality of the test voice. It is thought that this technique would overcome the rating scale judgment problem (Gerratt and Kreiman, 2001). As such, this measurement technique appears to support external validity in that most synthesizers (and arguably all current ones) permit users to adjust output qualities to their personal liking; such manipulation would almost certainly occur outside the laboratory in context use as well. Gerratt and Kreiman relate that listener agreement was significantly (and substantially) greater for the synthesis task than for the rating task, indicating listeners can in fact agree in their perceptual assessments of voice quality, and that analysis-synthesis can measure perception reliably (1996).

As mentioned, the MRT has been criticized for focusing primarily on the phonemes in the initial final positions. However, the focus on those elements is desirable when compared to, for example, the DRT's focus on only those errors that occur in the initial consonants. When comparing the MRT with phonetically-balanced word lists, the MRT is more desirable because it requires less training and is more quickly administered (due to the closed response set of the test). Given the current setting in which intelligibility is an issue (i.e., aircraft cockpits), one in which speech commands derive from a relatively simple vocabulary (i.e., one that is not highly variable with respect to content), it is argued that the MRT is more than sufficient as a measure of intelligibility. That is, the criticism levied against the MRT is one that is not an issue in the current context due to the non-complex nature of typical radio messages in aviation activities (e.g., an aircraft call sign with an airspeed change). As has been presented, the MRT has been used extensively in past studies pursuant to the evaluation of speech intelligibility. While other techniques may or may not be more externally valid, such as listener-

mediated analysis-synthesis, the wide acceptance and demonstrated validity of the MRT as an intelligibility measure make it a logical choice, at least until the other measures have undergone more empirical validation.

Human factors investigations of synthesized and digitized speech. It can be related that it was quite difficult to find *recent* research investigations with respect to the intelligibility of synthetic speech; indeed, most of the findings reported in the literature evaluating the perception of TTS synthesized speech were based on engines developed from 1979 to 1986. While a body of findings indicates reliable differences in comprehensibility levels between synthetic and natural speech, results vary considerably across different studies. However, several studies were located that are certainly germane and useful within the current context of aviation operations.

Using both a closed- and open-format MRT, Greene et al. (1986) tested eight TTS systems. The closed MRT condition provides information about phonemes that appear only in the initial and final positions; the open MRT provides information as a diagnostic aid in the identification of poorly synthesized phonemes by through an unbiased estimate of the most common types of perceptual confusions possible with each phoneme. Subjects were tested in groups of six in a quiet room containing individual cubicles, each equipped with a desk and a set of high-quality headphones. Subjects were presented with a single, isolated English word at each trial; their task was to indicate the word they heard on the answer sheet provided.

The results of the intelligibility tests indicate a wide range of performance for the different systems (see Table 5). The best performance for synthetic speech was obtained with the DECtalk Paul v1.8 in initial position (1.6%) and Prose v3.0 in final position

(4.3%). The worst performer was obtained with Echo for both initial and final positions (35.6% error rate for both conditions).

The researchers relate some conclusions based on the study. Four distinct groupings surface with respect to overall error rates: (1) natural speech, (2) high-quality synthetic speech (DECtalk Paul & Betty), (3) moderate-quality synthetic speech (Infovox SA 101, Berkeley, and TSI prototype I), and (4) low-quality synthetic speech (Votrax Type'n'Talk and Echo). The four groupings reflect the adequacy of the phonetic implementation rules used in the individual TTS systems which in turn is "directly related to the amount of speech knowledge incorporated into each system" (Greene et al. 1986, p. 105). The researchers further relate that these common error patterns across a wide range of synthetic voices suggests that some phonemes may be inherently difficult to perceive, especially since the phonemes typically misperceived in natural speech also tend to be those misperceived in synthetic speech. However, the error rates for synthetic speech were still substantially higher than those observed for natural speech. Further, the phonemes with the highest error rates were typically those with complex spectra or those showing the greatest amount of coarticulation in speech.

The Greene et al. research revealed a strong relationship between the amount of speech knowledge incorporated into a given system and the perceptual performance as measured by human observers. The gist of this conclusion is that 'one gets what one pays for.' High-end systems (such as DECtalk) have had the greatest amount of research and development and have been tested and evaluated more systematically prior to being offered to consumers (e.g., usability testing with focus groups).

TABLE 5

MRT error rates overall and error rates for consonants in initial and final position

	Error Rate (in percent)		
Voice	Initial	Final	Overall
Natural Speech	0.50	0.56	0.53
DECtalk 1.8, Paul	1.56	4.94	3.25
DECtalk 1.8, Betty	3.39	7.89	5.72
MITalk	4.61	9.39	7.00
Prose 2000 V3.0	7.11	4.33	5.72
Infovox SA 101	10.00	15.00	12.50
Berkeley	9.78	18.50	14.14
TSI-Prototype I	10.78	24.72	17.75
Votrax Type'n'Talk	32.56	22.33	27.44
Echo	35.56	35.56	35.56

Moreover, and perhaps more importantly, these systems include a more formal knowledge about the acoustic-phonetic properties of speech in the rule systems used to generate the synthetic speech.

Ricard and Meirs (1994) investigated speech localization and intelligibility from virtual directions, another study that has implications for cockpit auditory displays. Communication systems of modern aircraft typically carry two types of signals: speech from a variety of sources, and warnings (usually in the form of tones). Directional auditory cuing has been shown (Ricard and Meirs 1994; Begault 1993, 1998; Begault and Wenzel, 1992; Begault and Pittman, 1996) to reduce the time needed to locate a visual target. One use for directional filtering, then, is to add information about source location to an auditory display—information that was not there before—but another would be to increase the detectability of signals such as speech. Thus, the accuracy of direction estimates as well as the intelligibility of communication can be maximized as a design goal in systems that employ head-related transfer function (HRTFs). Head-related transfer functions are measured in an individual's ear canals; the data gleaned are used to

encode how sound waves interact with the human's hearing and characterizes how humans exploit signal propagation delays between the two ears to localize sound sources (Salvendy, 1997). Because subjects have shown variability in their localization of signals conditioned by *non*-individualized HRTFs, the Ricard and Meirs research sought to see if similar variability characterized the intelligibility of speech presented from synthesized azimuths (1994). In part, the researchers relate, this was to measure the gain (sensitivity) of speech intelligibility provided by directional filtering; they also wanted to see if anomalies of localization covaried with differences in intelligibility when both are measured within the same subject.

The measures were made with the MRT, and speech was produced and transmitted from a DECtalk v2.0 TTS system. In the experiment, the synthesized words were added to a continuous white noise that was band-limited to 0 Hz to 5 kHz with a roll-off of 96 dB per octave set to a spectrum level (i.e., the level of each individual frequency component of a signal) of 40 dB SPL. The speech and waveforms were led to separate channels of what is called a 'Convolvotron', where they were filtered according to azimuth, and then were presented on stereophonic headphones. Head position was measured with a Polhemus magnetic tracker (Ricard and Meirs, 1994).

The results indicated that subjects could accurately judge the direction of signals with simulated location information, especially when only differences in azimuth were present. Confusions of front and back present a difficulty for those who may attempt to apply directional sound technology. The rate of front/back confusions as well as the fact that their magnitude was greatest around the midline creates the challenge for an applied

technology of directional cuing. The researchers suggest that it may be better if virtual auditory displays used direction information as a redundant cue (Ricard and Meirs, 1994).

The aforementioned study has implications for attempts to provide three-dimensional (3-D) auditory localization in a SATS-like cockpit. This study, along with others that investigated warning tones in aircraft utilizing 3-D auditory displays (Begault, 1993 and 1998; Begault and Wenzel, 1992; Begault and Pittman, 1996), represents a relatively new applied human factors domain that shows promise. The Begault et al. studies have provided compelling evidence for the use of 3-D auditory displays in the localization of potential traffic conflicts both in the air and on the ground; the Ricard and Meirs study is the only research that could be found that investigated speech intelligibility in a 3-D auditory space (which could conceivably be extended to the cockpit). It appears that this arena is ripe for design initiatives that seek to capitalize on human audition characteristics for the presentation of information in the cockpit. Other investigations of a similar nature (especially those involving data link) will be presented below.

Tsimhoni, Green, and Lai (2001) studied the effects of natural and synthesized speech on driving performance. Using an IBM Embedded ViaVoice TTS Engine, 24 licensed drivers, equally divided by gender, drove a simulator on a road consisting of straight sections and constant radius curves, thus yielding two levels of low driving workload. The effects of message type (navigation, e-mail, news story) and voice type (TTS, natural human speech), and 'earcon cueing' (present, absent) were considered, creating a 3 X 2 X 2 within-subjects design. The control condition involved data collection while the participants were parked.

For all message types, the comprehension of the TTS messages, as determined by accuracy of response to questions, and by subjective ratings, was significantly worse than comprehension of natural speech. Driving workload was not found to affect comprehension. The researchers relate an interesting finding in that neither the speech format used (synthesized or natural) nor the message type (navigation, e-mail, news story) had a significant effect on basic driving performance, as measured by the standard deviations of lateral lane position and steering wheel angle. The results suggest that, in an operator performance condition, natural speech is superior to that of a TTS system in the comprehension of the message. The fact that performance was not affected is rather strange, in the opinion of the author, since the comprehension of the message directly relates to the performance resulting from that message. Especially when extended to potential SATS-like cockpit information displays, one could argue that miscomprehension of a message, either from ATC or aircraft operating in the vicinity, is a much more important finding, since incorrect or non-execution (i.e., slips or mistakes) of a maneuver has severe implications for air safety. The research finding above suggests the need for some kind of 'hyper-adjustable' digital speech system, one that can issue directives 'on the fly' through the use of some vast store of natural speech utterances that can be concatenated in real-time. Of course, this suggestion is likely beyond the limits of current technology, but with respect to aviation endeavors in which speech auditory displays are considered, such a system seems warranted.

Rehmann and Mogford (1996) investigated airborne data link. The FAA report resulting from that investigation suggested that pilots preferred digitized speech to a text-only presentation of messages on the system. The placement of the data link system was

also varied (i.e., below glare shield [center console], in between, or behind the pilots [aft]), with pilots preferring a center console position to an aft-mounted one. The availability of data link significantly reduced the amount of controller radio communication with 'pseudopilots' and simulator pilots. The subjective effort, workload of pilots, and fuel burn were not affected by the data link capability. However, pilots raised concerns about reduced confidence, safety, and situation awareness with data link.

The digitized speech preference over text really is not too surprising when one considers that the amount of 'head-down' time (i.e., scanning of instruments and displays, which requires a 'head-down' physical condition on the part of the pilot) is increased with textual systems in the cockpit, and that pilots often maintain their situation awareness through constant scans of both instruments and of the outside world through the windscreen. What would have been useful in the Rehmann study was if an additional independent variable had been introduced: synthesized speech. One could hypothesize that, based on previous studies (see above) of operator preference in demanding operational environments (such as driving, which has many similar elements to piloting) that the order of preference would be (most preferred to least preferred): natural (digitized) speech, synthesized speech, and textual format. This assertion exists as a focus of the current research.

The subjective results from participants in the Rehmann and Mogford study could simply be the result of the introduction of a new system to the cockpit. That is, new ways of performing ingrained (i.e., automatized) operations are typically met with concerns or fears of operational disruption and safety concerns. Once the system's interface is iterated through the techniques inherent in such fields as usability engineering, and is

demonstrated to be a valuable addition and its usefulness is established to the users, such concerns and resistance might disappear.

The above contention is further supported by the research investigations of Delgou et al. (1998), who studied the cognitive factors in the evaluation of synthetic speech. They showed that listening to and comprehending synthetic voices is more difficult than with a natural voice. However, and more germane to the argument presented previously, is that this difficulty can and does decrease with subjects' exposure to said voices. On the other hand, greater workload demands are associated with synthetic speech and subjects who listened to synthetic passages paid more attention than those listening to natural passages (Delgou et al., 1998). Perhaps repeated exposures to synthetic speech in the cockpit will follow a similar pattern—decreasing comprehension difficulties over time. This may have implications for pilot training, in that prospective pilots, perhaps, should be given synthetic speech media with which to listen to and become accustomed as their training progresses in an effort to place them on an even keel when initially exposed to cockpit systems utilizing synthetic speech. One could posit decreases in workload demands over time as well: perhaps these results are akin to the situation noted above—that comprehension and performance decrements are simply due to the 'newness' of the technology.

Paris et al. (2000) randomly assigned 78 participants in equal numbers to one of three speech modes: natural speech, DECtalk synthesizer, or Sound Blaster's Windows TTS synthesizer. The two TTS systems represent the current state-of-the-art with respect to the higher end (DECtalk) and lower-end (Sound Blaster) speech synthesizers. The researchers used the MRT for single-word intelligibility as well as an immediate-recall

task. Participants heard, in the MRT task, 50 words, and for each were asked to circle the word they heard from a set of six rhyming alternatives. Participants in the immediate-recall task heard 80 utterances, 20 of each type: normal (prosodic and contextual cues present), no prosody (normal sentences with prosody removed), no context (semantically anomalous sentences with prosody), and unstructured (unrelated words with no prosody), resulting in a 3x4 mixed design, with speech mode as the between variable and stimulus type as the within element.

The results for single-world intelligibility scores were as follows: 93.7% for natural speech, 83.1% for DECtalk, and 85.2% for Sound Blaster. This result is somewhat surprising in that the DECtalk is considered superior to and is considerably more expensive than the Sound Blaster product; however, the difference was not significant. The main effect of speech mode was significant, as was the main effect of stimulus type; there was also a significant interaction effect. Overall, participants judged natural speech to be the most intelligible, followed by DECtalk and Sound Blaster. For stimulus type, normal stimuli were found to be the most intelligible, followed by non-prosody stimuli, no-context stimuli, and unstructured stimuli. These results are outlined in Table 6.

The researchers conclude that, contrary to their expectations, the removal of prosody did not yield a performance reduction in the synthetic speech conditions. They proffer an explanation that the prosodic cues, such as those existing in speech systems, are not helpful, so their removal causes no performance decrement. Conversely, the prosodic cues evident in normal speech were very apparent, as performance deteriorated when they were removed.

TABLE 6

Comparison of DECtalk, Natural Speech, and SB Speech from Paris et al., (2000). Table (a) outlines the mean percentage of correct words as a function of speech mode and stimulus type; (b) outlines mean intelligibility ratings as a function of speech mode and stimulus type; (c) outlines mean naturalness ratings as a function of speech mode and stimulus type (p. 427).

(a)	Stimulus Type			
Speech Mode	Normal Sentence	No Prosody	No Context	Unstructured
Natural Speech	0.74	0.60	0.51	0.24
DECtalk	0.60	0.60	0.35	0.20
Sound Blaster	0.58	0.58	0.34	0.16

(b)	Stimulus Type			
Speech Mode	Normal Sentence	No Prosody	No Context	Unstructured
Natural Speech	9.86	7.81	9.27	5.90
DECtalk	8.20	7.74	6.13	6.07
Sound Blaster	7.10	6.39	5.22	4.11

(c)	Stimulus Type			
Speech Mode	Normal Sentence	No Prosody	No Context	Unstructured
Natural Speech	9.71	5.58	9.49	5.23
DECtalk	5.78	4.19	4.70	4.27
Sound Blaster	4.60	3.86	3.67	3.05

The researchers relate several design implications resulting from their investigation (Paris et al., 2000):

- *Prosodic cues*. Prosodic modeling as instantiated in the TTS synthesizers used in the present research does little to facilitate comprehension. This fact may explain why even high-quality synthetic speech still imposes a greater mental workload on listeners than does natural speech. Because performance is adversely affected, the use of synthetic voice in a task that requires rapid response to linguistic content, or in tasks involving linguistically complex or demanding secondary tasks, is questionable. This has implications for aircraft cockpit auditory displays: due to this increase in mental workload as a result of synthetic speech, airborne

data link systems that attempt to incorporate an auditory display may very likely have to utilize digitized speech; the addition of yet another mentally demanding task in this overcrowded workload environment does not appear to be justified.

- *Conceptual cues*. It is desirable for designers to incorporate as many contextual cues as possible within the limits of the specific task. Simpson and Williams (1980) recommend adding semantic context to synthetic cockpit warnings based on their findings that the additional linguistic redundancy provided by such cues reduced overall attention required for comprehension but did not increase response time. Further, context becomes increasingly important as intelligibility decreases, as in the high-ambient noise environs of a cockpit, wherein acoustical cues may be masked.
- *Comparison of TTS system quality*. Although single word intelligibility may provide some useful information, it does not assess differences that may exist in sequential prosody (i.e., phrases, sentences). Tests such as the MRT need to be supplemented with comparisons involving larger speech units. As such, TTS comparisons in the investigation of cockpit auditory displays need to incorporate such measures, especially when comparing the output of airborne data link, which will, more often than not, include longer passages relating vital information. Finally, designers should ensure that candidate TTS systems are capable of emulating the appropriate prosody.

Stern, Mullenix, Dyson, and Wilson (1999) investigated two TTS systems, the 'high-quality' DECtalk Express v2.4c, the 'low-quality' Monolouge DOS v1.1, and a

tape recording of human speech (digitized speech) in an effort to gauge the *persuasiveness* of synthetic speech and human speech. Their rationale for this particular research was that, since synthesized speech technology will soon be used in a variety of situations, investigations into the 'social factors' of their use are warranted. Put simply, the degree to which synthetic speech can be perceived as 'persuasive' is related as a worthy research endeavor. Further, TTS systems are perceived differently from human voice, and this arguably affects listeners' perceptions of the *speaker*.

One hundred ninety-three participants were randomly assigned to listen to an appeal under the three conditions mentioned above. The persuasive argument was a passage in favor of university-wide comprehensive exams that was adapted from models of strong arguments by Petty and Cacioppo in 1986. Default values for speech output were utilized in both TTS systems, and the default 'Paul' was used in the DEC product. Dependent measures were gleaned through questionnaires in which factors such as assessing speech characteristics, perceptions of the message, perceptions of the speaker, and the effectiveness of the message. The human speech condition involved five different speakers.

Results indicated significant differences between natural human speech and synthetic speech for six of the seven speech quality judgments that were measured. Human speech was perceived to be, as compared to synthetic speech, softer, higher pitched, less accented, less lengthy, less nasal, and livelier. The analysis of the speaker and message was conducted using a principal components analysis, which indicated five factors for the speaker: knowledgeable, truthful, powerful, involved, and accurate. Message factors included: captivating, clear, convincing, and simple. Human speakers

were seen as more knowledgeable, marginally more truthful ($p = 0.08$), more involved, and less powerful. Statistical contrasts examining differences between the two TTS synthesizers indicated that DECtalk was judged more knowledgeable and more involved than did the Monologue product. No significant differences were found between human speech and synthetic speech for factors such as attitudes toward the message or the effectiveness of the argument. Interestingly, and the main focus of the study, regardless of whether the message was listened to via human or TTS system, the message was found to be persuasive. Attitudes towards the message content, however, were significant. This suggests that, although the message was persuasive, the type of speech (human or synthesized) had no statistically differential effect on how persuasive the message was. The researchers conclude that most of the observed effects were due to the impoverished nature of synthetic speech produced by rule, which "leads the listener to view the computerized speaker as less knowledgeable, less truthful, and less involved"(Stern et al., 1994, p. 594).

The findings have direct implications for design issues. When TTS systems are utilized in social situations (related as interpersonal communication), in which personal attributes of the 'speaker' become important, the findings suggest that the differences that exist between TTS systems and natural speech may play a significant role in how people react to the user of a TTS system, such as users with disabilities. The researchers relate that evidence exists that the very use of technological assistance (such as TTS systems) by persons with disabilities affects others' perceptions of them (Stern et al., 1999).

Reynolds, Fucci, and Bond (1997) compared the effect of visual cuing on the intelligibility of DECtalk (version not specified) for native and nonnative speakers of

English in both ideal listening conditions and in the presence of background noise at a signal to noise (S/N) ratio of +10dB. The rationale behind the investigation is that, in the current climate in which improvements in micro processing and other technological abilities abound, it is imperative that speech synthesizers be intelligible enough to by easily understood by users, and not just those users who speak English natively, but those who speak it as a second language. The theme is that such insurance fosters intelligibility and supports usage by the increasingly culturally diverse population of the United States. The researchers relate previous research endeavors in which non-native speakers experienced significantly more difficulty in the transcription of sentences using the highly intelligible DECtalk system than did native English speakers. Thus, the current research question is whether there is any improvement in sentence transcription when a visual cue supplements the synthetic speech.

Twenty subjects each from native and non-native English speaking populations participated in the study. Thirty-two sentence pairs, which were randomly selected from an established inventory ("Sentences for Phonetic Inventory") were presented to each subject using a 'standard DECtalk male voice' (in this case, DECPaul). Half the sentences were presented in quiet and the other half in noise at a S/N ratio of +10dB. Background noise was introduced from a 'babble noise' tape from the SPIN test, which resulted from previous research. Half of the sentences were presented with visual cuing added to the first sentence of each pair and the other half were presented without cuing. Subjects read the sentences as it was being spoken by the DECtalk voice. Another sentence, topically related to the first, was then presented using only synthetic speech output; subjects were instructed to write down the second sentence of each pair

immediately after hearing it. The treatments were counterbalanced. This resulted in a mixed design; the between element was native or non-native, the within elements were environment (noise or without noise) and output (visual cuing or no visual cuing). The percent words correct in the second sentences was the dependent measure.

The results showed significant main effects for both group and environment as well as a significant interaction for group X environment. Visual cuing was not found to be significant for either group, although it *approached* significance for the non-native group (i.e., $p = 0.08$). The results suggest that visual cuing helps non-native speakers, but not in a manner that suggests it is wholly better than no visual cue.

Possible applications to the aviation domain include cockpit aids that seek to improve spoken (and thus transmitted) voice responses in very-high frequency (VHF) radio for non-native English aviators. While English is proscribed by the International Civil Aviation Organization (ICAO) as the international aviation language, it is known that, in foreign airspaces, pilots are often allowed to speak in that region's native tongue if they so desire (Illman, 1995). If visual aids were found to improve non-native users' understanding of English (which they were not) such visual aids might have been applied to cockpit environs in an effort to 'improve' the spoken English of aviators who speak barely-intelligible English (which happens quite often), simply as a tool with which to practice English skills while, for example, in straight-and-level international or oceanic flights in which there are long periods of relative inactivity. Then again, perhaps that would not be a good idea!

Lai, Wood, and Considine (2000) studied the effect of task condition on synthetic speech comprehension. 78 subjects were to evaluate the intelligibility of five *current*

TTS systems; however, and unfortunately for the current report, the researchers' goals were not to rank-order the tested systems—rather, they wished to understand if there were optimal conditions for synthesized speech comprehension, and to what degree comprehension might vary as conditions varied. As such, the conditions considered were the nature of the task and the effect of note taking while listening. The nature of the tasks was short, informal e-mail messages, longer messages, and 'CNN' news stories. The control condition was represented by comprehension measures of the tasks while read by a professional voice talent.

The five TTS engines used were DECtalk (v4.4), AcuVoice AV1700, IBM Via Voice Outloud (1998 version), L&H TTS engine v6.03, and Lucent Release 2. The results indicated no significant difference for comprehension performance of synthetic speech among the engines as evidenced by recognition memory. Additionally, although there was no subjective preference for male or female voice, subjects did perform better in the synthetic voice condition when listening to a TTS engine with a male voice (Lai et al., 2000).

With respect to the researchers' goals of the study, to determine the effect of task and note taking on comprehension performance, subjects were found to perform better with notes than without, with the medium and long passages fostering comprehension more than did the short passages (Lai et al., 2000). This finding may have implications for synthetic voice intelligibility in the cockpit, for while it is known that pilots often write down ATC instructions and information (e.g., heading commands, regional weather), the effect of synthetic speech of said information coupled with the fact that ATC transmissions (and those of local traffic) are by definition brief and to the point, as

well as the increased attentional demands that synthetic speech requires over natural speech (as evidenced by the research described above), use of the technology in aviation may do more harm than good.

Lee and Simpson (1998) conducted investigations of current and prospective voice warning systems in their RAPID (Rapid Pilot Interface Development Simulator) system that simulates the US Army's Apache Longbow assault helicopter. In their current configuration, the PVI (pilot-vehicle-interface) simulates voice alerts as warnings, caution, and/or certain feedback. A TTS synthesizer generates messages and the speech output is presented to pilots via headphones or a flight helmet. The researchers sought to determine the most effective (as evidenced through questionnaire) format of the spoken messages. A DECtalk TTS synthesizer (version unspecified), driven by a 'Smart Annunciation System' (unspecified) was utilized to generate the spoken messages. Comparisons were made between the current voice alerting system (digitized voice) and the 'current state-of-the-art,' the DECtalk system. Also investigated were different levels of information (full, terse, and clipped wording). The wording formats indicate differences in the amount of information that they relate. For example, 'full' messages provided the most information about a given threat; the 'terse' less so, and the 'clipped' provided little more information than a threat exists.

Interestingly, the researchers report that simple response time is not necessarily appropriate for the measurement of pilot's responses to tactical alerts. It is stated, "such measures do not take into account pilots' intentions or complex decision-making as they decide what to do about a particular alert" (Lee and Simpson 1998, p. 770). The researchers chose instead to rely on pilots' responses to questionnaires designed in

accordance with standard psychological rating scale design guidelines (e.g., utilizing likert or likert-type scales). One wonders whether usage of questionnaires in evaluation of candidate auditory messaging systems in GA aircraft is also suggested instead of or in addition to response time.

Pilots were found to be 'extremely satisfied' with the new voice type afforded by the DECtalk TTS system and judged it as more intelligible than the existing one. Additionally, pilots desired the ability to control the level of detail (i.e., full vs. terse) provided by the voice about the threats; that is, they wanted the ability to 'declutter' the voice at pilot discretion. Pilots did prefer, however, the 'full' version as it provides as much information as is available about a given condition. Since these are automated systems, such ability is not feasible in GA operations, as actual humans provide the signals. Several pilots noted that they could glean time-critical information more readily from the voice without having to go 'eyes inside' (i.e., head-down) to the visual display.

Further results suggest that TTS systems are more versatile in future iterations of warning systems because not only can they handle all alerts (i.e., tactical, system, and flight parameter), but they provide the capacity to handle known *new* demands for alerts through providing a cost effective solution to growth (Lee and Simpson, 1998). Further, the TTS system, in comparison to the current digitized words and phrases, allows more flexibility and growth potential. Finally, the researchers state, "the challenge will be to design the voice messages so that they behave like a good co-pilot and provide this detailed information without saturating the pilots' auditory capacity" (Lee and Simpson 1998, p. 772).

With the increasing levels of automation envisioned for next-generation GA operations, the requirements for monitoring, processing, and response are intensified, as pilots will increasingly be 'on their own,' for the most part, with respect to traditional ATC tasks of traffic avoidance and station-keeping. As such, in vehicles with advanced transport systems, speech technologies are continually being investigated as a means of providing critical route and navigation information while decreasing mental workload and improving safety. With the use of auditory displays in the cockpit come issues of optimal presentation levels, especially when one considers the requirement of simultaneous performance of the visual (i.e., scanning both the instruments and the outside environment) and manual tasks (i.e., actuation of various control surfaces as necessary) in an environment that (typically) is dominated by low-frequency engine noise. This leads to concerns of optimal presentation intensity of auditory displays, and this design element has been demonstrated to be a key factor affecting the detectability, compliance and perceived urgency of non-verbal warnings (Baldwin and Struckman-Johnson, 2002). Momtahan (1990) found loudness level to be one of the acoustic parameters most significantly associated with the perceived urgency of non-verbal warnings, with louder sounds generally having been judged as more urgent than less-intense sounds. Other parameters affecting perceived urgency included 'inter-pulse interval length', 'spectral shape', and 'number of harmonics'. Methods have been established for the determination of the appropriate loudness level for non-verbal auditory warnings and for the appropriateness of other key factors in intensity (Edworthy, 1994). Although many research endeavors have investigated numerous aspects of auditory speech processing, speech intensity research has focused mainly on detectability.

Also, as Baldwin et al. (2002) point out, researchers frequently do not report the decibel (dB) level used in the presentation of speech stimuli in their experiments, thereby disallowing ease of comparison across studies. Indeed, as presented in the Baldwin research, since intensity level has been reported to affect perceived urgency of non-verbal warnings, it may very well impact verbal warnings as well.

Speaks, Karmen, and Benitez (1967) have examined the presentation levels associated with optimum speech intelligibility in environments with low background noise. They found that the percentage of correct identification of sentences within a quiet background rose sharply between presentation intensities of 20-30 dB. Detectability and intelligibility are essential in understanding auditory speech processing, yet both falls short of quantifying the amount of mental effort required to process the stimuli. As reported by Haas and Casali (1995) in actual operational environments, listeners are frequently performing several simultaneous tasks and are thus unable to devote their complete attention to auditory tasking. As Baldwin (2002) points out, in a multi-task situation, quantification of the cognitive resources required by a more difficult task (e.g., traffic avoidance) would leave the pilot with fewer spare resources with which to allocate to any additional tasks (e.g., listening to ATC commands within Class B airspace) that had to be performed simultaneously.

With the advent and iteration of active noise reduction (ANR) headsets, however, one wonders whether their use within the cockpit would constitute 'low background noise' and if their use might mitigate these effects. Indeed, their use might prove moot arguments for optimal presentation levels as related to ambient noise in the cockpit, for such noise is 'spectrally cancelled' via the ANR circuitry. Further, they may have the

potential for improving the safety and performance of all pilots, and may even be essential for older pilots experiencing presbycusis effects, although regulating bodies (e.g,. ANSI, ISO) stop short of labeling ANR devices as 'hearing protection.'

Rehmann (1996, 1997) has conducted research investigations of data link systems that utilize digitized speech and/or textual formats within commercial aviation operations. A hypothesis of the former study was that a digitized announcement of incoming data link messages would improve pilot response time and result in reduced head-down time, and evaluated three message presentation formats: radio, data link text format, and data link text format plus digitized speech. The provision of digitized speech was thought to obviate the need for the pilot flying (PF) to glance at the data link display unit. Rehmann (1996) found data link WILCO response times differed significantly between text-only and text/digitized speech modes, with the time required to respond increased with digitized speech, which at first appears odd. However, this was suggested to be the result of the *cadence* of the digitized speech—pilots typically could read the text message before the digitized speech completed its utterance, and WILCO actuation could only occur once the message was finished. Indeed, Rehmann relates it is likely that reading text from the screen will always be faster than hearing it read aloud. As a result, according to Rehmann, it may be necessary to institute a cockpit procedure to WILCO first so that controllers receive an early indication that the aircraft intends to comply with an ATC instruction. Time spent in reading the *full* message aloud (as opposed to an abbreviated one) was found to be significantly reduced when using digitized speech. This result appeared to relate to a reduction in the need for crew coordination (i.e., that text required a full, verbatim reading and that digitized speech did not resulting in

increased 'discussion' amongst the crew). Interestingly, subjective evaluations indicated that data link was seen by pilots as promoting less confidence and perception of safety. When asked about their preferences for digitized speech over a text-only presentation, pilot written comments were in favor of the digitized speech presentation, although some improvements were recommended. Rehmann concludes (1996) by suggesting the speed of speech be increased and it should not interfere with other radio traffic.

In-vehicle investigations using auditory displays. Future iterations of the NAS include aircraft that utilize state-of-the-art glass cockpit displays that are envisioned to portray not only the traditional 'six-pack' (i.e., the six most oft-used instruments: airspeed indicator, artificial horizon, vertical speed indicator, turn indicator, heading indicator, altimeter), but also HITS and other informational items. As more visual displays are added to aircraft, not only does the magnitude of visual information processing increase, but the requirement to shift attention between different visual displays also increases.

Early investigations of synthesized voice within the confines of the cockpit have demonstrated measurable performance benefits with their use. Simpson and Williams (1980) found that, during the most visually, manually, and cognitively demanding approaches in simulated commercial operations, performance with synthesized voice was superior to that of the normal procedure of pilot-not-flying verbal callouts. Even though their experiment involved commercial operations utilizing pilot 'teams' (i.e., 'pilot flying' and 'pilot not flying') one wonders whether such results would replicate within single-pilot GA operations. Indeed, with respect to recent advances in both speech

processing capability and avionics coupled with envisioned NAS architectures, this appears to be a ripe research question.

In driving tasks, which are argued to possess many similarities with piloting (e.g., maintenance of station keeping, monitoring of traffic location and variability), visual attention switching has been linked to decremental performance, especially in older drivers (Baldwin and Schieber, 1995). These effects are not limited by age—Hagar and Payne (1996) found that attention switching was detrimental to their participants' abilities to perform concurrent tasks; again, by extension, the 'aviate, navigate, and communicate' triumvirate of piloting operations most certainly can be considered concurrent tasks.

Auditory displays can be superior to visual displays in the presentation of navigation and warning information, but the literature appears mixed. In simulator studies evaluating in-vehicle navigation devices, Walker, Alicandri, Sedney, and Roberts (1991) found that drivers using auditory navigation devices of varying complexity made significantly fewer navigation-related errors than those of using visual mode devices. In addition, in high driving workload situations, drivers using auditory displays did not reduce their speeds as much as those using visual devices did. However, when auditory displays are compared with multimodality displays, the effects are even murkier. Liu (2001) conducted a driving study concerning 'ATIS-like' auditory information (i.e., similar to aviation ATIS but specific to driving) in the form of a digitized female voice using a SoundBlaster PC soundboard, and incorporated both a multimodality and visual display. Under high driving load conditions, participants tended to drive faster when using the auditory display alone than with either the visual or the multimodality display. Further, visually presented complex information resulted in poorer vehicle control, as

evidenced through more frequent lane deviations, than with either the auditory or mulitmodality displays. The auditory display condition resulted in the lowest workload rating, even lower than that of the multimodality display (Liu, 2001). Baldwin and Struck-Johnson (2002), in their driving tasks supported with an auditory speech display, utilized what at first appeared to be a different dependent measure—the time to complete the track (i.e., a driving course). Upon further dissection, however, their measure does indeed mirror, however indirectly, those presented above; the time to complete the track is wholly dependent on the speed at which the participants navigated the task.

It has long been understood that operators respond faster to voice warnings than to visual ones (Simpson, McCauley, Roland, Ruth, and Williges 1987; Sorkin 1987). Though the proposed research outlined herein is not specifically involved with warnings per se, voice warnings *do* occur within piloting operations (e.g., TCAS; conflict warnings from ATC). Further, the current trend of traffic location and maintenance depiction on state-of-the-art visual displays suggests other concerns; such heavily loaded visual displays have been shown inferior to auditory displays with respect to time-sharing performance (Wickens, Sandry, and Vidulich, 1983). Indeed, for safe driving, short auditory information coupled with visual display may optimize perceptual and cognitive performance. Liu (2001) hypothesized that the improved results obtained through multimodality display in his research may be due to smaller attentional demands than either of the single display modalities, and the workload results of that study supported this contention. One could posit that, in GA operations, this optimization may be mirrored; for example, short auditory ATC messages specific to local traffic position are supported through a visual traffic display that can be referenced. However, due to the

nature of flying, which requires constant monitoring 'out the window' (i.e., 'head-up' or 'eyes inside'), at least in visual flight rules (VFR) and approach conditions, the increased head-down time related to any visual element may *increase* workload. Indeed, the auditory display condition of Liu's research (2001) resulted in the lowest workload rating, even lower than that of the multimodality display. In short, it is unclear whether such effects on workload can be transferred to the aviation domain with respect to indices such as airspeed or in-trail station keeping maintenance; this is another question the current research sought to address.

Situation awareness (discussed in detail later) is a very important component of motor vehicle operation. Good driver SA can be said to consist of knowledge about the environment, road geometry, weather and its effects on visibility, traffic information (e.g., vehicle configurations, rate of flow), and driver behavior (e.g., own and others' intentions). It will be discussed how SA plays a role in the safe operation of *all* human-operated vehicles, especially aircraft.

Deatherage (1972) defined a set of guidelines for the selection of auditory or visual display channels based on characteristics of the message, the environment, and the task. The guidelines state that auditory presentations are indicated when the messages are short, simple, and temporal in nature; require immediate action; and do not have to be referred to later. ATC commands may meet these message guidelines, as they are, by necessity, short; they may not be as simple as the layperson would define it, but, within the highly-specialized environment of piloting, in which the commands are usually the same in construct and largely vary in numerical content only, they can be said to be simple. These messages most certainly require immediate action due not only to the

speed in which aircraft operations occur, but also to the varying traffic densities that exist depending on location. As ATC routing commands require timely performance response, they have little need to be referred to later, and thus meet the last of Deatherage's requirements. Conversely, other ATC messages (e.g., weather information) might not meet these recommendations.

The workload associated with aircraft displays depends on the complexity of several items, not the least of which are spoken commands from ATC; the interaction requirements necessary to manipulate the radio system (i.e., location of the radio microphone [if handheld]), as well as the time pressures usually associated with them (i.e., pilots cannot usually wait until they are 'free' to respond as ATC requires quick, succinct, and correct responses to their commands), can increase workload dramatically. These requirements place attentional demands on operators that often result in dual-task processing during instances of high workload, such as within traffic patterns or operations in class B airspace.

Which speech synthesizer to use? The literature seems to agree that the most effective speech synthesis systems available are DECtalk systems or those that utilize its technology (although the last study mentioned did suggest the effectiveness of other systems). Indeed, even in investigations that occurred in the mid-80s, wherein early prototypes of DEC-powered systems were explored, the DECtalk systems were superior, especially over the then-standard VOTRAX systems, which, by all indications, were subjectively horrid and relatively unintelligible.

As technology progressed, other systems became available, but even within the past several years (i.e., 1997-2000), DECtalk systems appear to be consistently superior.

As mentioned, DECTalk systems have had the greatest amount of research and development and have been tested and evaluated more systematically prior to being offered to consumers (e.g., usability testing with focus groups), so it is not too surprising that it performs better than other systems. Moreover, and perhaps more importantly, these systems include more formal knowledge about the acoustic-phonetic properties of speech in the rule systems used to generate the synthetic speech. In many cases, the DEC-powered systems were shown to be as good as natural speech within certain treatment conditions, and were often found to be superior to anything else tested with respect to intelligibility. These results are supported in the literature reviews discussed above.

Especially when one considers speech synthesis applications for data link, the capabilities of the DEC systems are almost a requirement. In 1993, for example, there were 255 near midair collisions that were the direct result of communication errors between pilot(s) and ATC; this value represents 15% of all near midair collisions for that year (Prinzo, 1996). These sobering findings suggest a need for speech synthesis systems that are as close to approximating natural human voice as can be applied. However, this suggestion must be tempered with other results indicating that the very use of synthesized speech causes decrements in reaction time due to the increased attentional requirements associated with synthetic voice perception. Any investigations of speech synthesis for the cockpit will therefore need to carefully consider these results. If one were to investigate speech synthesis systems for aviation operations, including in the support of SATS-like operations, the DEC-powered systems such as DECtalk are suggested. However, newer speech synthesizers have surfaced in recent years that have not been

evaluated empirically for intelligibility and, at least subjectively some newer systems (e.g., AT&T's Natural Voices) sound more realistic, and may thus approach natural voice more successfully than DECtalk systems. It is therefore of interest to compare one or more of these newer systems against DECtalk, which has been extensively studied, to see if perhaps a newer system may be indicated for use in GA operations. Such an evaluation is presented and discussed later.

With respect to digitized systems, and as noted, virtually any system that possesses the capability of digital playback (via digital audio tape [DAT] or other media) can effectively function as a digitized speech system. It should be noted that there were no studies found comparing these systems, so data as to superior-performing particular makes or models cannot be related. There do exist, however, and as discussed above, several available digitized speech systems, and many are marketed as augmentative devices and/or have the capability of recording digital input as well. As such, suggested digitized speech systems for future GA investigations include DAT tape recorders/playback devices, which are made by several vendors (Akai, Korg, Roland, Tascam, Yamaha), but can range greatly in price (from $300 to over $5000). Alternatively, one can opt for one of the many specialized digitized speech systems noted previously, for they are just as capable.

Situation Awareness

History of SA. The Air Force Tactical Command once stated that the difference between a good fighter pilot and a dead fighter pilot is situation awareness (Gawron, 2002). The dramatic growth of situation awareness has been fostered by many factors,

chief among which are the challenges posed by new classes of technology. Tools used in complex systems with which to aid humans in the performance of tasks have focused not only on physical tasks, but with the rather elaborate perceptual and cognitive tasks as well. Pilots operating in the complex airspace proposed by future airspace operations such as the SATS must be able to sufficiently perceive and comprehend a huge array of data, which almost by definition will be highly dynamic. The growth and complexity of electronic systems and automation, especially those that may be required for the SATS, have driven designers to seek new methodological frameworks and tools for effectively dealing with these changes. Additionally, one must understand that technological systems do not inherently provide SA: it is the human operator who must usefully apply perceived information to satisfactorily reach system goals. Endsley (2000) relates that a large gap exists between this deluge of data produced and presented to the pilot (through whatever modality) and the pilot's ability to filter that data such that only germane informational bits are utilized for decision making; this has been termed the 'information gap'. Systemic sensors, such as traffic displays and other instruments, collect some subset of all available information from the system's environment and internal system parameters, and some portion of this is displayed to the operator via its interface. Of this information, the pilot perceives and interprets some portion, resulting in SA. However, one must take caution to not assert that more data equals more information. Indeed, the implementation of automation in the cockpit simply for the sake of automation has been shown to exacerbate this problem by not taking into account these tenets (Endsley, 2000). It is therefore a question of providing the pilot with needed information in such a way that it is useful both perceptually and cognitively. The emphasis of SA in current system

design has occurred for two reasons: (a) because designers can now do more to ensure that good SA is provided through the implementation of decision aids and system interfaces, and (b) designers are concomitantly able to actually hinder these same efforts if we fail to adequately address the SA needs of pilots (Endsley, 2000).

Endsley promotes a practical example of what embodies the tenets of SA and its successful implementation: the Gulf War. The war, occurring in the early 1990's, was said to be the first 'information war' (Endsley, 1997). The Coalition forces sped up their focus in collecting, disseminating, and using information in an effort to successfully produce new tasking orders within seventy-two hours, rather than the current (at that time) temporal dislocation of several weeks. The Iraqi's flow of information, by contrast, was severely disrupted through coalition bombing runs that destroyed command and control centers (C^3) and power grids for communication systems.

It has been related (Gawron, 2002) that the US Army has maintained the same hierarchy of forces (i.e., corps, division, brigade, battalion, and company) since the time of Napoleon. The advent and progression of information technology and its application within the private sector has caused organizations to 'flatten' (i.e., no longer as hierarchical as before), and has widened spans of control (i.e., made operations more 'horizontal' or 'lateral') because everyone can (ideally) have the same situation awareness of where that particular organization is and where it is going (Gawron, 2002).

The advantages of SA are many. SA permits the seizure of the initiative early, be it on the battlefield or within the GA context. In military operations, SA reduces the enemy's reaction time; for quick, succinct, and correct information can ideally be acted upon before the adversary even realizes that they are compromised. SA permits more

mental energy to be applied to current and future situations, leading to decreased time spent on 'housekeeping' (i.e., maintaining stores, mundane communications); SA permits the timely and accurate use of all resources. Good SA increases both the speed of planning and execution of a goal, and increases the efficiency and the effectiveness of that goal (Maggart and Hubal, 1998).

SA defined. SA is defined in terms of the goals and decision tasks relevant to a particular operational environment. That is, SA will differ based on application. The pilot has no need to know every detail of his/her immediate environment (e.g., the type of sunglasses worn by a passenger) but does have an obvious need for knowledge related to the goal of safe operation of the aircraft. SA has traditionally evolved from the specific domain of aviation, wherein the term was coined. The earliest definition that could be found was that provided by Melanson, Curry, Howell, and Connelly (1973, p. 70):

> Knowledge of (the pilot's) current position with respect to the air route structure, knowledge of the position of other aircraft around him, the ability to predict evolution of the traffic situation, and the ability to choose an appropriate escape route in an emergency.

Endsley has iterated this definition and economized it somewhat to define SA as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" (2000, p. 3).

As described by Roscoe, Corl, and LaRoche (1997), SA is the ability to:

- Attend to multiple information sources,
- Evaluate alternatives and establish priorities,

- Estimate probable outcomes for different courses of action,
- Work on whatever has the highest momentary urgency without losing sight of the routine,
- Record priorities as situations deteriorate or improve, and
- Act decisively in the face of indecision by others.

Further, specific to the aviation domain, Garner and Assenmacher (1996, p. 147) state that SA is:

> Staying ahead of the other aircraft, knowing what's going on so you can figure out what to do, detecting information in the environment, processing the information with relevant knowledge to create a mental picture of the current situation, and acting on this picture to make a decision or explore further.

SA therefore requires knowledge of both the internal and external states of the humans and systems, the system/environment relationship, and the environment itself (e.g., temperature, position, terrain). Successful and appropriate attainment of this knowledge is typically explained through Endsley's 'SA Levels.'

SA levels. SA is comprised of three (3) levels. Level 1 SA is simply the perception (or detection) of cues (indeed, one cannot begin to assimilate data and form a correct picture of the operating environment unless it is perceived). Jones and Endsley (1996) found that 76% of SA errors in pilots could be traced to perceptual problems related to needed information due to either failure of the system or cognitive shortcomings. Level 2 SA is centered on the ability to adequately comprehend what is perceived; that is, identification. This level is associated with the ability to successfully filter the myriad data being perceived in terms of their relation to operational goals.

Endsley and Garland (2000, p. 4) proclaim that it is the components of this particular SA level that sets SA apart from earlier psychological research and places it firmly in the realm of "ecological validity". Level 3 SA is the highest level of SA, comprising the ability to project situation elements and dynamics into likely future occurrences, or prediction of what is going to happen based on successful attainment and application of knowledge and information acquired within the previous levels. This level represents the mark of a skilled expert in the domain of interest (Endsley and Garland, 2000).

The ability to forecast from current events such that future events are anticipated fosters timely decision-making and is a definite boon to aviators, as the time and speed of aircraft operations necessitate a requirement for this quality in an effort to avoid conflicts. Smith and Hancock (1995, p. 138) have defined SA as an "adaptive, externally directed consciousness", and take the position that SA is a purposeful behavior that is goal-directed in a specific task environment. They have also proposed another definition (1995, p. 138) as relates to cognition; SA is "up-to-the minute comprehension of task relevant information that enables appropriate decision making under stress". Other definitions include Sarter and Woods (1991, p. 46), who state that SA is the "accessibility of a comprehensive and coherent situation representation which is continuously being updated in accordance with the results of recurrent situation assessments". Lave (1988, p. 124) further states that SA "fashions behavior in anticipation of the task-specific consequences of alternative actions". Furthermore, different researchers work toward differing practical ends and these ends affect how SA is defined. In pilot selection, for example, SA is defined as a 'talent', whereas pilot training requires an SA definition as an 'improvable skill' (Metalis, 1993).

Whatever definition one chooses, a pilot who has SA is akin to an 'expert' who can look at a huge array of discrete stimuli and immediately integrate them into 'chunks' or meaningful bytes of knowledge upon which he/she can base appropriate action. An expert pilot sees the view outside and the cockpit instruments and perceives the human/aircraft system flying with respect to relevant others through space and time. But, unlike other experts, who may focus their attention on only one topic, the pilot must be able to multitask between several different subsystems, and must do so not at a personal pace but within the time and priority constraints dictated by the flying environment (Metalis, 1993). Endsley and Garland (2000) caution against confusion of the term with *situation assessment*, which is defined separately from SA in that it is an active process of seeking information from the environment, and that SA is the result of that process. Finally, Endsley (2000) proffers that there is no such thing as 'too much SA'; more is always better. Indeed, the simplest operational definition of SA is that it is that information that one 'really needs to know.'

The temporal aspect of the SA definition, the 'within space and time' element, relates to the fact that operators constrain parts of the world (or situation) that are of interest to them based not only on space (how far away that element is) but also on *how soon* that element will have an impact on the operator's goals and tasks. This has implications for envisioned predictive displays (e.g., HITS or traffic predictions). The dynamic nature of airborne situations dictates that the situation is always changing, so the pilot's situation awareness must constantly change (or be rendered 'outdated') and is thus inaccurate (Endsley, 2000). This forces the operator to adapt many mediational (cognitive) strategies for the maintenance of SA. The role of others in the process of SA

development must also be considered. Verbal and non-verbal communication with others (including radio communications, hand signals and 'wing tipping' of pilots) has historically been found to be an important source of SA information. Even in situations with restricted visual cues, ATC report that they get a great deal of information from just the voice qualities of the pilot's radio communications, deriving information on experience levels, stress, familiarity with English instructions, level of understanding of clearances and need for assistance (Endsley, 2000). This has implications for research investigations that attempt to integrate synthesized and/or digitized speech interfaces in SATS-like environs, for a valuable SA tool may be modified such that these cues are no longer useful or usable.

In the current context, the question of what is to be evaluated is understood to be potential cockpit and ATC systems that maintain current standards of safety while simultaneously supporting the increased capacity that a future GA system likely requires. One notable element in the current and future NAS is the development and maintenance of shared mental models of traffic—shared in the concept of pilots and ATC as team members—which can be said to be a much more difficult task when team members are distributed in terms of space, time, and/or physical barriers, as one could argue the case to be in the context of SATS operations. This includes, but is not limited to (Endsley, 2000):

- *Shared situation awareness requirements*: the degree to which the team members know which information needs to be shared, including their higher level assessments and projections, and information on team members' task status and current capabilities;

- *Shared SA devices*: the devices available for sharing this information, which can include direct communication, both verbal (in the sense of VHF radio traffic from both ATC to pilots and vice versa, as well as from aircraft to aircraft in the local area) and non-verbal (e.g., wing-tipping), shared displays or a shared environment. As non-verbal elements in a shared environment are usually not available in distributed teams (i.e., pilots and ATC), this places more emphasis on verbal communication and technologies for creating shared information displays;
- *Shared SA mechanisms*: the degree to which team members possess mechanisms, such as shared mental models that support their ability to interpret information in the same way and make accurate projections regarding each other's actions. The possession of shared mental models can greatly facilitate communication and coordination.
- *Shared SA processes*: the degree to which team members engage in effective processes for sharing SA information, which has been found to include a group norm of checking assumptions, checking each other for conflicting information or perceptions, ensuring coordination and prioritization of tasks, and establishing contingency planning, among other processes.

One could argue that the concept of shared mental models can be assured at best and supported at worst through the use or provision of shared displays. Such displays could foster communication, via whatever modality, but need to be thoroughly examined

from a system standpoint. This need is supported through flight simulation experiments supporting envisioned NAS operations (Endsley, 2000).

***Decision making, memory, and attention*.** Decision-making is a separate and distinct process from SA. Indeed, SA is represented as the main precursor to decision making. Endsley presents several reasons for this. First, it is entirely within the realm of possibility for a pilot to have perfect, ideal SA yet make an incorrect decision. Endsley (1995) found that 27% of aircraft accidents involved situations wherein there was poor decision making even though the aircrew appeared to have adequate SA for decisions. On the other hand, it is possible, through sheer luck, to make correct decisions when SA is not optimal or is poor. Decisions are formed by SA, and SA is formed by decisions: they are inextricably linked (Endsley and Garland, 2000). However, according to the researchers, SA is not decision-making and decision-making is not SA. Perhaps a question of semantics, this distinction has implications for the measurement of SA. Several factors have influences with respect to the accuracy and completeness of SA that an individual pilot derives from the environment. The way in which attention is employed in this highly complex arena with multiple competing cues is an essential element with which to determine what aspects of the situation will be processed to form SA. Once attended, this information must be integrated with other information, it must be compared to goal states, and it must be projected into the future—all of which are heavily demanding on working memory (Endsley and Garland, 2000). In this condition, both the perceptual salience of environmental cues and the meaningful direction of attention of the pilot are important. Indeed, the correct prioritization of information in this dynamic environment remains a challenging aspect of SA. Endsley and Garland

(2000) relate investigations reporting four strategies utilized by operators in an effort to reduce the working memory load associated with SA, including the aforementioned information prioritization, chunking, 'gistification' of information (i.e., encoding only relative values of information where possible), and the restructuring of the environment to provide external memory cues. Endsley and Garland (2000) further demonstrated that pilots could report on relevant SA information for five to six minutes following freezes in an aircraft simulation without the memory decay that would be expected from information stored in working memory. This result was hypothesized to support a cognitive model that suggests working memory is an activated subset of long-term memory (LTM) (Endsley and Garland, 2000). Put in this sense, SA can be said to be a unique product of external information acquired, working memory processes and the internal LTM stores activated and brought to bear on the formation of the internal representation. Further, Endsley (1988, 1995) hypothesized that LTM stores play a major role in dealing with the limitations of working memory.

Perceptual cues may come in the form of visual, aural, tactile, olfactory, and taste receptors. In the aviation domain, pilots and ATCS are able to directly view and hear information from the environment itself. The concept is illustrated in Figure 2.

In an investigation of airborne data link, Rehmann (1996) found interesting results with respect to SA. Data link was generally well received by flight crews and, as mentioned earlier, crew subjective effort (i.e., workload) was not affected by the presence of data link capabilities. Pilot concerns mainly focused on SA issues. There were clear indications of a loss of awareness for navigational information regarding surrounding aircraft when using maps and probe questions. As mentioned, the effects of SA reduction

on pilot performance are not yet fully established, especially within GA operations. The commercial pilots within the Rehmann study voiced their concerns about confidence, safety, and SA reduction in general. However, and following one of Rehmann's hypotheses, SA was found to decrease overall for the data link flights (over radio-only), especially for information about other aircraft (1996).

Rehmann concludes by suggesting that methods need to be developed to offset these SA decrements when data link is in almost exclusive use as a communication medium, as it is sure to be within future iterations of the NAS.

SA requirements analysis in GA. The design of interfaces that provide and support SA depends upon domain-specifics that determine the features of the situation that are relevant to a pilot. Typically, three methods are utilized in the specification of SA requirements:

- *Method 1*: specification of all information that is needed
- *Method 2*: specification of all information being used by observing current systems
- *Method 3*: specification of all information that is needed using digital models

With respect to the first method, the focus is on identifying and providing for all categories of SA information (i.e., geographical, spatial/temporal, system, environmental, and tactical). To that end, Endsley (1999) proposed use of a goal-directed task analysis for the determination of SA requirements.

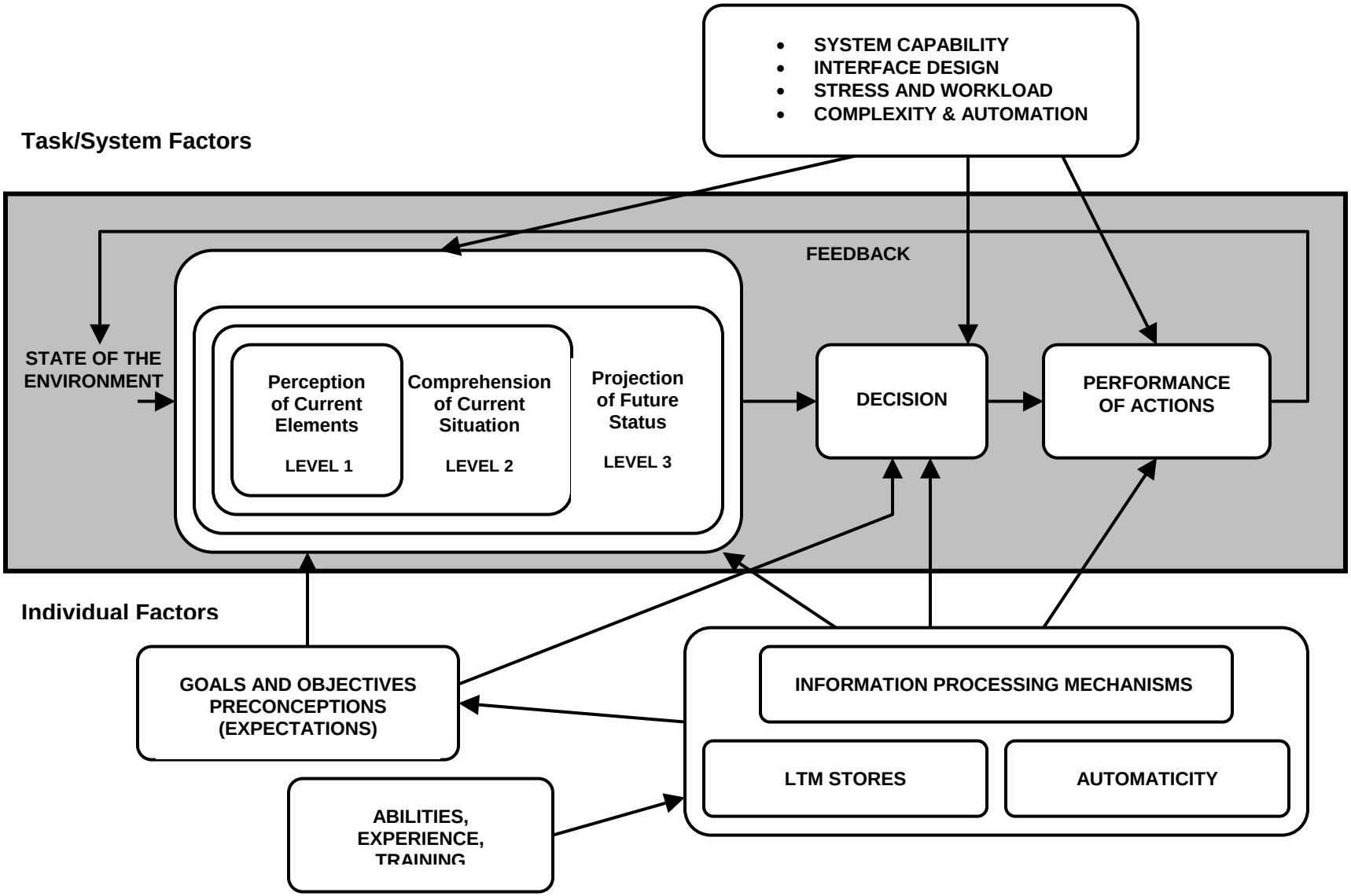


Figure 2. Model of SA in dynamic decision-making. From Endsley & Garland (2000), p. 3.

The methodology focuses on basic operational goals and the SA requirements necessary for each decision (see Figure 3). The requirements are stratified with respect to the three aforementioned SA levels (basic data, integration and comprehension, and future projection).

Goal
Subgoal
Decision
Projection (SA Level 3)
Comprehension (SA Level 2)
Data (SA Level 1)

Figure 3. Format of goal-directed task analysis in GA.

The result provides information on not only what information to supply, but also how it needs to be integrated to support operational SA. Endsley (1999) relates a combination of cognitive engineering procedures with which to glean this information, such as expert elicitation, observance, verbal protocols, analysis of written materials and documentation, and the use of formal questionnaires. The data is then pooled and validated by a larger number of operators. The process differs from traditional task analysis in that: (1) it is not set to a fixed timeline (which is not compatible with dynamic flight environments), (2) it is technology independent, is not tied to how tasks are performed but to what information is ideally needed, and (3) the focus is not only on what is needed, but how that data is integrated to support decision making and goal attainment (Endsley, 1999). Endsley (1997, pp. 3-4) provides several SA requirements that are applicable across many aircraft systems:

- *Geographical SA:* Location of own aircraft, other aircraft, terrain features, airports, cities, waypoints and navigation fixes; position relative to designated features; runway & taxiway assignments; path to desired locations; climb/descent points;
- *Spatial/Temporal SA:* Attitude, altitude, heading, velocity, vertical velocity, G's, flight path; deviation from flight path and clearances; aircraft capabilities; projected flight path; projected landing time;
- *System SA:* System status, functioning and settings; settings of radio, altimeter and transponder equipment; ATC communications present; deviations from correct settings; flight modes and automation entries and settings; impact of malfunctions/system degrades and settings on performance and flight safety; fuel; time and distance available on fuel;
- *Environmental SA:* Weather formations (area and altitudes affected and movement); temperature, icing, ceilings, clouds, fog, sun, visibility, turbulence, winds, microbursts; instrument flight rules (IFR) vs. VFR conditions; areas and altitudes to avoid; flight safety; projected weather conditions;
- *Tactical SA:* Identification, tactical status, type, capabilities, location and flight dynamics of other aircraft; own capabilities in relation to other aircraft; aircraft detections.

An example is described with respect to the second method. Consider a Fire Commander who must collect, describe, and analyze a fire scene (Martin and Flin, 1997). His subordinates, a Station Officer (SO) and the Assistant Divisional Officer (ADO)

provide the data (see Figure 4). As can be seen from the Figure 4, the most frequently reported information to the Fire Commander allows him/her to create data that can be analyzed for relevant SA information. For example, results from Figure 4 can be mapped to strategic and planning information (e.g., communications), information about resources utilized (e.g., number of fire engines and equipment), the type of fire (e.g., heat, smoke), location (e.g., town, type of building), people involved (e.g., resource requirements, evacuation), and investigation aspects (e.g., cause). This information can provide powerful information with which to analyze current firefighting activities such that they can be optimized.

Another example is presented with respect to the third method of SA requirement formulation. During an infantry SA workshop in 1998 at Fort Benning, SA requirements were investigated for infantry combatants and teams (Gawron, 2002). Four relevant aspects of activities were expertly analyzed germane to: (1) soldiers, (2) platoons, companies, and battalions, (3) brigades, and (4) future elements.

The expert analyzers functioned as a 'human consultant system' to evaluate several indices during the workshop:

- Commander's intent (two up and one down the command chain)
- Succession of command (in the event of simulated 'death')
- Environmental data (ground, weather)
- Coalition/reserve/interservice/civilian visibility data
- Enroute updates, inter-aircraft links, mission rehearsal for troops in motion
- Individual soldier status
- Enemy and location of troops

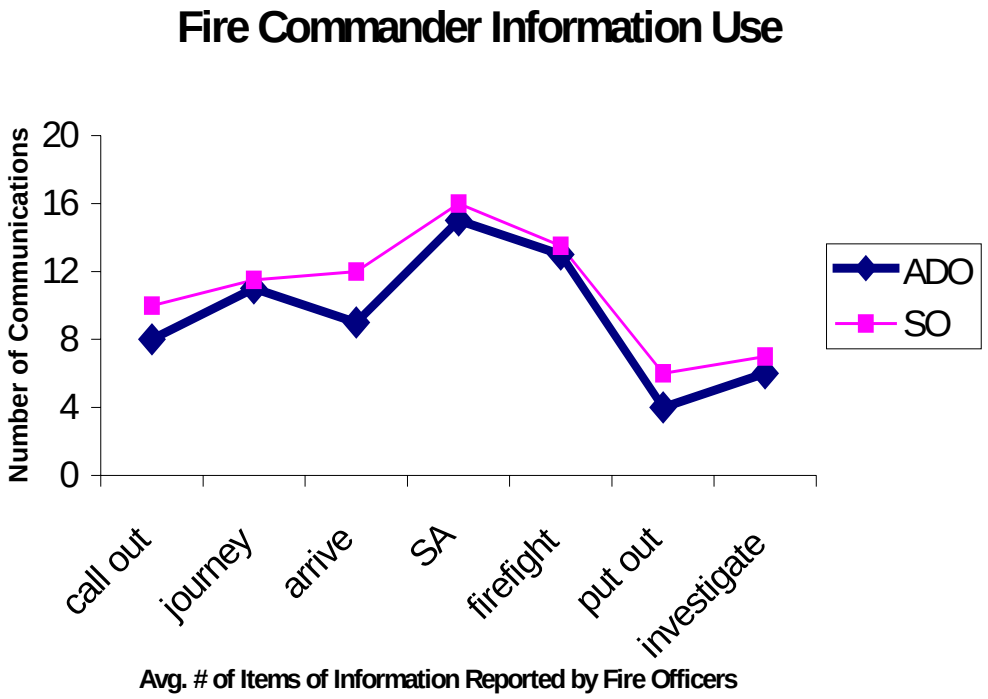


Figure 4. Example of SA requirement method 2 (from Martin and Flin, 1997, p.2/4).

Situation awareness is widely recognized as a critical element within aviation operations. However, almost all of the research to date has focused on military or commercial transport pilots who are typically highly experienced; GA pilots, on the other hand, being generally much less experienced, are considered much more prone to aviation accidents, and the data support this contention. General aviation accidents account for 94% of all US civil aviation accidents and 92% of all fatalities in civil aviation through July 1999 (Trollip and Jensen, 1991). Even though GA accident statistics are generally good, GA pilots continue to have mishaps due to several pilot-related factors. Trollip and Jensen (1991) report that the pilot was found to be a "broad cause/factor" in 84% of all GA accidents and 91% of all fatal accidents. A substantial percentage of GA accidents were declared related to poor decision-making and, as a

result, it appears evident that GA operations lack a clear understanding of SA requirements.

Shook, Bandiero, Coello, Garland, and Endsley (2000) sought to provide some data in this regard and conducted an investigation into situation awareness problems within GA. Among their findings:

- Landing and approach phases are the most problematic, followed by take-off, taxi-out, and climb phases;
- Student pilots working toward their instrument rating were found to have the least SA problems overall;
- Multi-engine pilots were more frequently rated as having moderate to frequent problems with SA across most phases of flight; and
- Problems with SA were found to significantly decrease with experience across most phases of flight.

The researchers relate several key problem areas that need addressing with respect GA, including focusing on task management, basic procedures, vigilance, awareness and effects of weather, dealing with malfunctions, building mental models, and critical skill development (Shook et al., 2000). SA requirements focus not only on what the operator needs, but also on how that information is integrated or combined to address each decision that is made. SA requirements are defined as those dynamic information needs associated with the major goals or sub-goals of the operator in performing his/her job (Endsley and Garland, 2000).

The aforementioned goal directed task analysis seeks to determine what operators would ideally like to know to meet each goal. Based on these results, observations made by the author from previous GA experience and experiments, and including private pilot interviews, an attempt was made to perform an SA requirements analysis specific to the current GA experiment utilizing 'method 1' (specification of all that is needed); see Figure 5.

Rehmann (1993) conducted an SA requirements analysis for commercial operations that might be affected by data link (see Table 7). Even though for commercial operations, several elements of the Rehmann analysis are germane to GA as GA operations most certainly will continue within the terminal airspaces typically serviced by commercial operations. As such, several items within this list are candidates for inclusion as SA probes.

As Endsley and Garland relate, the SA requirements can be "whole, or for just particular goals or subgoals of the operator" (2000, p. 4). As such, the requirements list in Figure 4 contains only those items that are germane to the current experiment. The SA requirements form the basis with which to determine the SA queries for use within the domain of interest (i.e., GA vectoring operations within a class C airspace).

Endsley further states that the determination of which queries should be provided should be based on three things: 1) SA requirements analysis, 2) the capabilities and limitations of the simulation and simulation scenarios, and 3) the objectives of the test (Endsley and Garland, 2000).

Goal 1. Assess Safety

1.1 Assure aircraft (a/c) is operating within safety limits (Supports system SA)

1.1a Critical powerplant operations within range?

- What are RPM, oil pressure/temperature/mixture levels?
- Engine RPM, oil pressure/temperature/mixture currently at safe level?
- Engine RPM, oil pressure/temperature trending normally (if applicable)?

1.1b Control surfaces position?

- What portion of flight am I in?
- What is my flap/landing gear/trim position?
- Is current flap/landing gear/trim position ideal?

1.2 Assure a/c position is safe (Supports geographical and spatial/temporal SA)

1.2a Aircraft altitude?

- Minimum safe distance from terrain?
- Minimum safe distance from obstacles?
- Minimum safe distance from other aircraft?

1.2b Aircraft attitude?

- Is a/c straight-and-level, climbing, or descending?
- Is a/c turning?
- Do I need to change my attitude to maintain safe operation?

1.2c Aircraft airspeed?

- What is the indicated airspeed (IAS)?
- Is IAS safe w.r.t. control surface position?
- Do I need to change my airspeed to maintain safe operation?

Goal 2. Assess Communication and Compliance with ATC

2.1 Correct Radio Setting for local ATC? (Supports system SA)

2.1a Am I on the correct frequency?

- What is the frequency for local tower?
- Is radio stack set to local tower frequency?
- What frequency do I need to know next?

2.1b Understand directive(s)?

- What was the last ATC communication?
- How does the last ATC communication relate to my current status and to those of other a/c around me?
- What will a/c position be in relation to ROA at completion?

2.2 Conformance with ATC Directives? (Supports spatial/temporal SA)

2.2a Acuating to assigned altitude?

- What is assigned altitude?
- What is my current deviation from assigned altitude?
- Will I be ascending/descending to assigned altitude?

2.2b Acuating to assigned airspeed?

- What is assigned airspeed?
- What is my current deviation from assigned altitude?
- What is my groundspeed?

2.2c Acuating to assigned heading?

- What is assigned heading?
- What is my current deviation from assigned heading?
- How long until I reach assigned heading?

Figure 5. SA requirements analysis for GA operations.

Since the SA requirements analysis has been 'focused' to operations specific to the current experiment and taking into account the capabilities/limitations of the simulator, the final concern is to ensure inclusion of queries that support the experimental objectives; that is, the evaluation of data link formats.

Endsley and Garland (2000) caution, however, to not focus the queries so narrowly that only one or two items of interest are visited because subjects will likely shift their attentions to these few items and therefore artificially inflate SA. Thus, with respect to the current experiment, it is important to ensure that the queries are not limited solely to verbiage and indices related through ATC.

TABLE 7

Situation awareness items and sources for items identified by Rehmann (1993, p. 18). Items marked as 'yes' are candidates for SA probes within GA operations.

SITUATION AWARENESS ITEM	SOURCE	AFFECTED?
NEXT COMM. FREQUENCY	ARRIVAL/FINAL ATC	YES
WEATHER SITUATION	ATIS	NO
WEATHER SITUATION	AIRCRAFT	YES
TRAFFIC SITUATION	ARRIVAL/FINAL ATC	YES
SEQUENCING	ARRIVAL/FINAL ATC	YES
HOLD SITUATION	ARRIVAL/FINAL ATC	NO
TERMINAL ROUTING	ARRIVAL/FINAL ATC	YES
APPROACH CLEARANCE	ARRIVAL ATC	YES
CONTROLLER ERRORS	ARRIVAL/FINAL ATC	YES
MISSED APPROACH	TOWER ATC	NO
WINDSHEAR	AIRCRAFT	NO
GO AROUND	AIRCRAFT/TOWER ATC	NO
AIRCRAFT ON RUNWAY	AIRCRAFT/TOWER ATC	NO
BRAKING ACTION	AIRCRAFT/TOWER ATC	NO
TAXIWAY TURNOFF	TOWER ATC	NO

The formats of ATC directives will differ only in the presentation utilized (e.g., synthesized, digitized, textual)—not that of content. This is because ATC transmissions are comprised from a relatively small vocabulary—typically numerical information such as in altitude, airspeed, and heading—all three of which may be specific to the receiving

aircraft or to local traffic; weather advisories (as appropriate), and frequency changes round out the list. There are several verbal statements or phrases also utilized by ATC in combination with the numbers, for example, '*turn left to*', '*climb and maintain*', '*state position*', '*squawk 2600*.' To account for and plan against artificial inflation of SA, then, the current experiment also sought to include other queries that are not specific to ATC directives; for example, *fuel remaining*, *flap setting, radio setting*, *RPM*, etc. Based on the SA requirements analysis, the Rehmann (1993) work, and with respect to the considerations outlined above, the following SA queries were selected for use in the current experiment (not a complete list):

1. What is your current airspeed?
2. What is your aircraft altitude?
3. What was the last ATC communication?
4. What is your aircraft heading?
5. What is the make/model of the last traffic advisory?
6. What is your aircraft's assigned runway?
7. What was the position of the aircraft during the last traffic advisory?
8. What was the last known location of any traffic?
9. What was the altitude directive of the last ATC communication?
10. What was the altimeter setting of the last ATIS message?
11. What is your current deviation from your intended/assigned heading?
12. What are the weather conditions at the airport as related by ATC?
13. What is the trajectory of the last traffic advisory relative to ownship?

Another question with the use of SA queries is *how often* should subjects be probed? Jones and Endsley (2000) found that the number of events (i.e., real-time probes, SAGAT, and/or secondary measures) should be increased from that used in their study (one every two minutes) in an effort to increase sensitivity. Subjects did not find the events intrusive as long as they did not occur during verbal communications. As such, the current experiment introduced probes at least once every two minutes. The ordering of the particular queries was counterbalanced taking into account the current

phase of flight (i.e., a query about a particular ATC directive will not come before the directive is presented).

Mental models and SA. Long term memory (LTM) stores in the form of mental models. Critical cues in the environment can be matched with internal schemas to indicate prototypical situations that provide instant classification of situations and comprehension (Endsley and Garland, 2000). Scripts of the proper actions to take may be connected to these prototypes, thereby simplifying decision-making. The use of mental models in achieving SA is considered to depend on the ability of the pilot to 'pattern match' between critical cues in the environment and elements of the mental model. In this respect, SA is the current state of the mental model (Endsley and Garland, 2000).

The concept of a mental model, as related by Endsley and Garland, is useful in that it provides a mechanism for: (a) guiding attention to situation-relevant aspects, (b) an avenue with which to integrate perceived information to form an understanding, and (c) a means for projecting future states based on current states and understanding of its dynamic nature (2000). Without the mental model, the integration of data and its projection would be prohibitively difficult, yet experts (e.g., line pilots within commercial operations) appear to be able to perform these tasks with ease. Visual scanning may be assumed to be driven by a mental model of the process whose elements are being displayed. Indeed, Bellenkes, Wickens, and Kramer (1999) relate that breakdown in scan is one of the leading contributors to mishaps where loss of situation awareness (LSA) was identified as a causal factor. The expert pilot's mental model of

flight dynamics, which drives the scan across the instrument panel, is complex, reflecting the complexity of the dynamics themselves.

There are three features that make these dynamics particularly challenging: first, attention is limited and therefore to some extent the pilot must trade-off the allocation of resources between the three primary tasks or axes of control (i.e., longitudinal, vertical, lateral axes). Appropriate allocation of resources to axes that require positive control (because they are changing), while not altogether neglecting those that must be monitored so they don't diverge from target values, requires a high skill of attentional flexibility (Bellenkes et al., 1997). Second, all three axes are somewhat sluggish, defining the traditional higher order effects (i.e., second and, in the case of lateral deviations, third order), which presents a need for the consultation of predictive displays, such as the vertical speed indicator. Third, dynamics are interactive in complex ways. For instance, an increase in bank causes a decrease in pitch and thus an airspeed increase. With experience, however, pilots develop internal models of the systems they operate and the environments in which they operate. Information is extracted more efficiently by experts from nearly all instruments, and particularly from the high bandwidth, information-rich attitude directional indicator.

In general, there is no standardized method of teaching tactical scanning, and instructors are normally unable to confirm whether the pilot is actually scanning effectively (Bellenkes et al., 1999). This may have implications for experienced and inexperienced pilots attempting to aviate within future NAS iterations; the former may have trouble iterating long-held, established models of operation (due to the arguable 'paradigm shift' in operational activities); conversely, the latter may display accelerated

adaptation to SATS-like operations resulting from the 'malleable' state of their internal representation of flight operations. However, this suggestion is not supported in at least one investigation (Lancaster et al., 2003), in which experienced pilots performed superior to inexperienced pilots in their ability to maintain glide slopes that were increasingly deviant from established convention. Perhaps this result was due to experts possessing more automatized skill in extracting information, and their performance is a result of a more refined mental model.

Goals and errors in SA and their relation to aviation. Pilots typically will have multiple goals that may shift in importance as a flight progresses. The goals direct the selection of the mental model, which will serve to direct attention in information selection from the environment. For this reason, selection of the correct goal is an extremely critical aspect in attaining ideal SA (Endsley and Garland, 2000). If the incorrect goal is pursued, critical operational elements may be overlooked and may lead to false comprehension of the environment. Endsley and Garland (2000) relate that this process is a top-down, goal-driven process in which the goals actively guide information selection. A simultaneous bottom-up process occurs in which informational elements are utilized to iterate SA, and a given situation assessment can lead to choice of a new goal. So while a pilot may be engaged in the goal of navigation, the chiming of (for example) the pitot freeze alarm will (hopefully) trigger the implementation of a new goal, which directs selection of a new mental model and focus of attention in the environment (Endsley and Garland, 2000).

At any level, an error in SA can be induced by deficiencies in system design (i.e., the needed information is not available, is poorly presented, is ambiguous, or is in the

incorrect format), or by information processing errors (memory or attentional limitations, pattern matching failures or in mental projection) (Endsley, 1999). Endsley (1999) notes that SA errors are rarely independent; that is, a failure at attending cannot always be placed on the operator, but rather is also a function of system design. By gaining an understanding of why SA problems occur, it may then be possible to design systems that account for these deficiencies.

Endsley (1999) suggests that there exists three levels of SA error as well as a few others (see Table 8). At the most basic level, errors may be the result of inadequate perception. While it is true that some information may not be available to the pilot due to a system deficiency, some data is indeed available but, for various reasons, is not utilized, either due to exclusion from the scanning pattern (omission), external distractions, or attentional narrowing (tunneling). Endsley further states (1999) that this 'missing of available information' was the single largest causal factor for SA errors (31.5% in a study of commercial carriers using NTSB data).

Level 2 failures include instances where information is correctly perceived but its significance or meaning is not comprehended. This error class may be the result of lacking a good mental model for the combining of information in association with applicable goals. In other cases, the wrong mental model may be utilized in information interpretation. For example, the mental model of a similar system is used to interpret information, leading to a false diagnosis or understanding of the situation in areas where the system differs (Endsley, 1999).

Level 3 failures exist in the highest (skill) level of SA. Pilots may be fully cognizant of activities in their airspace, but are unable to adequately project what that

means for the future. This may be due to a poor mental model or some other reason. Since Level 3 is at such a high level and is thus a very mentally demanding task, it isn't too surprising that failures occur at this level, although it is a relief that these errors accounted for a small percentage of accidents (3.4%) in the study (Endsley, 1999).

TABLE 8

SA error taxonomy (from Endsley, 1999, p. 3).

Level 1: Failure to correctly perceive information
- Data not available
- Data hard to discriminate or detect
- Failure to monitor or observe data
- Misperception of data
- Memory loss

Level 2: Failure to correctly integrate or comprehend information
- Lack of or poor mental model
- Use of incorrect mental model
- Over-reliance on default values
- Other

Level 3: Failure to project future actions or state of the system
- Lack of or poor mental model
- Over-projection of current trends
- Other

General
- Failure to maintain multiple goals
- Habitual schema

The other types of errors in the table represent the two general categories of causal factors. Some pilots have been found to be poor at multiple goal maintenance, which could impact SA at all three levels. Additionally, evidence exists that people fall into a "trap of executing habitual schema, doing tasks automatically, which render them less receptive to important environmental cues" (Endsley, 1999, p. 5).

SA Measurement. Endsley and Garland (2000, p. 17) suggest that "SA is a beneficial concept precisely because we can measure it," and that it "provides a great insight into how operators piece together the vast array of available information to form a coherent operational picture." SA measures further provide a valuable index with which to evaluate system design and for an improved understanding of human cognition. The researchers further relate that, as an 'intervening variable' in between stimulus and response, SA measures provide far greater sensitivity (ability to distinguish changes) and diagnosticity (indications of variation as well as the cause of that variation) than is typically available for performance measures. One of the main reasons for measuring SA is for evaluating new system and interface designs (such as cockpit auditory display of ATC information). It is necessary to systematically evaluate new technologies and design concepts as relates to their improvement (or decrement) of pilot SA; this provides evidence with which to base design decisions. Further, the explicit measurement of SA determines the degree to which a design objective has been met. SA can be directly measured along with mental workload and performance measures. Endsley and Garland (2000, p. 17) state that:

> High level performance measures are often not sufficiently granular or diagnostic of differences in systems designs, and, while one system design concept may be superior to another in providing the pilot with needed information in a format that is easier to assimilate, the benefits of this may go unnoticed during the limited conditions of simulation testing or due to extra effort on the part of operators to compensate for a design concept's deficiencies.

For this reason, direct measures of SA will foster selection of design concepts that promote SA and thus provide a means with which pilots can make effective decisions and avoid ineffective ones. Undesirable elements such as data overload, non-integrated data, automation, and complex systems that are not easily understood as well as many other factors can be identified early in the design process and corrective changes can be made to improve the design (Endsley and Garland, 2000). As is known within the human factors community, such early intervention is precisely the correct avenue to take in order to avoid costly and often time-consuming redesigns further into the development cycle, and any tool that enhances the human factors engineer's ability to do this is most welcome indeed. Such a tool would be useful in the investigation of data link display modalities in a SATS-like environment…

To address these goals adequately, the veracity of available SA measures must be established. The measure must be valid (it measures the construct that it claims to measure) as well as reliable (can repeatedly result in the same conclusion) in addition to the diagnosticity and sensitivity qualities noted above. Different pilots may utilize different processes with which to glean data (information acquisition methods) to arrive at the same knowledge state, or they may arrive at different knowledge states based on the same processes due to differences in comprehension and/or projection. As such, SA measures that tap into SA processes may provide information as to *how* a pilot has reached his/her informational state.

However, Endsley and Garland (2000) caution that such measures will only provide 'partial and indirect' information with respect to a pilot's *level* of SA. Further, while there may be an experimental need for such information, care should be exercised

in any attempt to infer one from the other. Endsley and Garland (2000) posit that the relation between situation awareness and performance can be viewed as a 'probabilistic link'. That is, good SA should increase the likelihood of good decisions and performance, but does not guarantee it, and the opposite may be true. However, it is related that poor performance does not, in many cases, result in serious error (e.g., disorientation at low altitude is much more likely to result in an accident than at high altitude).

As such, SA measures can be said to only indirectly represent behavior and performance. Further, Endsley (1997) relates that measures of workload only capture half of the picture: how hard the person is working—not what benefit they are gaining for their efforts. It is imperative that an SA measurement technique does not intrude on the pilot's attentional distribution, as this may well change the construct that is being measured. Direct measures of SA, according to Endsley and Garland (2000), tap into a person's knowledge of dynamic environmental state. Such information may reside in working memory or LTM to some degree under differing circumstances. A significant issue related by Endsley is that attempts to tap into memory may affect the degree to which operators can report mental processes to make such information accessible (1997). Additionally, temporal aspects, as already noted, may affect an operator's ability to report information from memory. As is known, with time there is a rapid decay of information in working memory; only LTM access may be available. Research has shown (Nisbett and Wilson, 1977) that recall of mental processes after the fact tends to be over-generalized and over-summarized and rationalized, and may thus present an inaccurate view of SA processed dynamically. On the other hand, real-time access of information

from memory can also be problematic in that such access may influence ongoing performance, decision processes, and SA itself. Real-time access may affect information gleaned through various modalities as well, since it is known, for example, that auditory stimuli are ephemeral and cannot be referred to as can visual stimuli, which are often much more static; these associations must be carefully considered when employing any SA measure that attempts to determine the appropriateness of candidate SATS-like displays.

Each class of measures for SA may have certain advantages and disadvantages in terms of the degree to which a given measure provides an index of SA, as well as their possible intrusiveness for use in in-flight SA assessment in simulation. Additionally, the objectives of the researcher and any experimental constraints will have an impact on the appropriateness of a given measure of SA. A discussion of the relative merits and liabilities of various SA measures is described below.

China Lake SA (CLSA). Developed by and for US military pilot training, the China Lake situation awareness is a measurement technique requiring operators to provide a rating of 1 through 5 either during or after a flight of their SA (Gawron, 2002). Table 9 diagrams the CLSA scale.

The CLSA technique is strong in that is maintains high face validity (i.e., it appears to be a valid measure to those that use it), it has clear content definitions, it fits into flight cards, and is relatively easy to administer. However, it is somewhat limited in that it is a subjective rating, has seen limited use (specifically to operational flight tests only), ratings can only be made during 'benign' portions of flight, and, arguably most

importantly, it is not yet validated. However, with continued use, especially within the military, one can see its value in the future (Gawron, 2002).

Crew SA. Another technique for SA measurement within the cockpit is that of 'Crew SA' (Gawron, Weingarten, Hughes, and Adams, 1999). Within this SA measure, expert observers are utilized to rate crew coordination. Usually, this is accomplished through one of two methods: (1) the observer is physically present, typically sitting behind the crew in a 'jump seat', or (2) the observer measures and catalogues information post-facto through videotape analysis.

TABLE 9

China Lake situation awareness scale.

SA SCALE VALUE	CONTENT
VERY GOOD **1**	• Full knowledge of A/C energy state/tactical environment/mission; • Full ability to anticipate/accommodate trends
GOOD **2**	• Full knowledge of A/C energy state/tactical environment/mission; • Partial ability to anticipate/accommodate trends; • No task shedding
ADEQUATE **3**	• Full knowledge of A/C energy state/tactical environment/mission; • Saturated ability to anticipate/accommodate trends; • Some shedding of minor tasks
POOR **4**	• Fair knowledge of A/C energy state/tactical environment/mission; • Saturated ability to anticipate/accommodate trends; • Shedding of all minor tasks as well as many not essential to flight safety/mission effectiveness
VERY POOR **5**	• Minimal knowledge of A/C energy state/tactical environment/mission; • Oversaturated ability to anticipate/accommodate trends; • Shedding of all tasks not absolutely essential to flight safety/mission effectiveness

The expert observer develops what are called 'transfer matrices' which foster classification of 'decision' or 'non-decision' information. Use of Crew SA has had mixed results. It is clearly strong in that it is sensitive to the types of errors that can occur: minor, moderately severe, and major (operationally significant) errors. Further, it is sensitive to decision prompts; that is, when an occurrence presents that requires an immediate decision (e.g., turn right to avoid a potential conflict). However, it is somewhat limited in that it requires open and frequent communication among crewmembers. It can be difficult (as in usability testing) to require operators to verbalize their thoughts and decisions; indeed, the requirement for verbalization can be said to disallow normal operations (e.g., crews might not usually fully articulate what is happening and may rely on gestures). Additionally, and more relevant to the current research, is that Crew SA requires a team of expert observers (Gawron, 2002).

Snapshots. 'Snapshots' is another SA technique that is primarily used within military circles. It requires expert observers to select appropriate 'points in time' within a particular training regimen, wherein trainees state the status of own and enemy forces. The expert observers take this data and compare the actual and perceived status of those enemy forces (Gawron, 2002).

Again, as this is another military-themed SA measure requiring expert observers, it is neither appropriate nor indicated for use in real-time simulations of advanced GA concepts. Additionally, it requires some time to complete the evaluation of the observations—it is not typically immediate. However, it does retain strength in that it is an objective measure that is applicable to any system; indeed, it has been utilized with

success within commercial airline training in the evaluation of electronic taxi chart displays (Amar, Hansmann, Hannon, Vaneck, and Ghaudhry, 1995).

Situation Awareness Global Assessment Technique (SAGAT). If SA is to be a design objective, then it is imperative that it be specifically evaluated during the design process of candidate SATS-like displays. Failure to do this will result in inability to determine if proposed concepts actually support SA, do not support SA, or inadvertently compromise it in some way (Endsley, 1999). The SAGAT (Endsley, 1988b) has been successfully applied in the aviation domain, in display design, and in interface technologies. This knowledge elicitation technique is currently being used to evaluate everything from graphic displays for aircraft, to automation concepts to advanced free flight to the F-22 Raptor tactical fighter (Endsley, 1997).

SAGAT provides an objective measure of SA based on queries during freezes in a simulation. The freezes are periodic and randomly timed, and incorporate a 'blank' of all operational displays during the freeze. At the time of the freeze, a series of queries are provided to the operator in an effort to assess his/her knowledge of what was happening at the time of the freeze. The queries are determined based on the previously discussed (for GA operations) in-depth task and requirement analyses, which must be conducted for each domain in which SAGAT is used. Operator responses to the queries are scored based on what was actually happening in the simulation at the time of the freeze (within operationally determined tolerance zones)(Endsley, 1998). The score, then, is the operational definition of SA (Metalis, 1993).

The main advantage of SAGAT is that it provides an objective, unbiased index of SA that assesses operator SA across a wide range of indices that are germane for SA in a

given system. The main disadvantage, however, is that it requires freezes in the simulation. Because the freezes are random and cover such a broad spectrum of operator SA requirements (see above), operators cannot prepare for the queries and it has been found (Endsley, 1995) that the freezes do *not* affect performance in simulations. SAGAT has been criticized for its reliance on memory, but, as mentioned previously, studies have shown that this real-time access is superior to 'recalling after the fact' because the latter is often over-generalized and over-summarized and/or rationalized. However, and as discussed above, real-time queries such as those used in SAGAT can be affected by working memory limitations. This does not appear to be a problem because reviews of the literature (Dreyfus, 1981; Nisbett and Wilson, 1977) suggest that such problems are indeed a concern when operators are asked to report *how they know* something, and *not* what their assessments of the situation are. The queries typically last from 2 to 5 minutes, depending, of course, on the number of queries provided (Endsley, 1998). Additionally, Endsley (1995) found that subjects were able to effectively report their assessments for as long as 5 to 6 minutes during SAGAT freezes without memory decay being a problem. This result indicates that the SA of experienced operators performing tasks in a system with high ecological validity (i.e., real task domains and not artificial laboratory tasks) is accessible for verbal report via a "fairly stable internal representation"(Endsley 1998, p. 2). Further, since the queries are highly salient since they are direct extensions of the SA requirements analysis, they maintain high content validity. SAGAT has been found to possess another desirable quality, predictive validity (prediction of operator performance), at least in an air combat task (Endsley, 1998).

Other strengths of SAGAT include that fact that it is (as mentioned) an objective measure, which is highly desirable with respect to replication. It is applicable to any complex system, and maintains empirical, (as mentioned) predictive, and content validity (Gawron, 2002). For SAGAT to be utilized successfully it requires real-time, human-in-the-loop simulation and it must contain appropriate queries that are germane to the current context. These qualities suggest SAGAT would be very useful for in-flight investigations of the kinds of auditory displays that are attractive in future aviation systems.

Situation Awareness Rating Technique (SART). SART provides a subjective rating of SA by operators in systems (Taylor, 1990). This technique has a total of 10 components, which were determined through analyses with pilots to be relevant to SA. Pilots rate on a series of bipolar scales the degree to which they perceive (1) a demand on resources, (2) supply on resources, and (3) understanding of the situation (Endsley, 1998); see Table 10. The scores are then combined to provide an overall 'SART score' for the system. SART ratings have been found to be correlated with operator performance in evaluations in cockpit designs (Selcon and Taylor, 1990), so are certainly germane to studies with the i-GATE, and with subjective measures of workload (Selcon, Taylor, and Koritsas, 1991).

The main advantages of SART is that it is easy to use and can be administered in a wide range of tasks. Further, it does not require any customization for differing domains and can be used in real-world investigation as well as simulations (Endsley, 1998). SART is sensitive to different types of decision-making, is easily administered,

and it is sensitive to tasks involving aircraft attitude recovery and learning comprehension (Gawron, 2002).

However, Endsley (1995) cautions of disadvantages to SART use, specifically subjective concerns: (1) the inability of operators to rate their own SA (i.e., not knowing what they don't know or what errors may exist in their own mental models), (2) the possible influence of performance on their ratings (i.e., operators may provide ratings based on how well they *think* they are doing), and (3) possible confounding with workload issues (i.e., attentional focusing).

TABLE 10

SART Rating Scale

		LOW						**HIGH**
		1	**2**	**3**	**4**	**5**	**6**	**7**
DEMAND	Instability of Situation							
	Variability of Situation							
	Complexity of Situation							
SUPPLY	Arousal							
	Spare Mental Capacity							
	Concentration							
	Division of Attention							
UNDER	Information Quantity							
	Information Quality							
	Familiarity							

Additionally, SA may operate as an independent factor from workload in many situations (Endsley, 1993). However, it has been posited (Selcon et al., 1991) that the combination of SA and workload factors into one scale may provide parsimony in the process of data collection. Criticisms of SART state that, although it is cost-effective and can be used in simulation and in real flight, evaluations must rest on the somewhat dubious assumption that the evaluator is consciously aware of all the mental elements of what constitutes SA (Metalis, 1993). Finally, it is ordinal data (rank order), and therefore is not a candidate for rigorous statistical evaluation, yielding nothing quantitative about the differences between the scale's levels.

SA Subjective Workload Dominance (SA SWORD) Technique. SA SWORD (Vidulich and Hughes, 1991) is a technique requiring subjects to complete a rating scale that lists all possible pair-wise comparisons of tasks performed. Pair-wise comparisons are such that, for example, subjects compare observations '1 & 2', '1 & 3', '1 & 4', '2 & 3', '2 & 4', etc. An evaluator constructs and completes 'judgment matrices' constructed from these ratings. The ratings are then calculated using 'geometric matrices' (unspecified).

Strengths of SA SWORD include such desirables as sensitivity to differences in tracking tasks and any color displays that might be used, and it enjoys a test-retest reliability of +0.937. Criticisms of its use are that it is insensitive to the use of any flashing for cueing (if indeed used) and that it requires everything to be designed such that pair-wise comparisons are formulated, and it requires software for rating calculation (Gawron, 2002).

SA Linked Instances Adapted to Novel Tasks (SALIANT). SALIANT (Muniz, Stout, Bowers, and Salas, 1993) is a very complex, yet comprehensive, SA measure. It requires *extensive* data collection, usually over an extended time period. SALIANT is designed around five 'phases' of data collection. The first phase involves observation and cataloguing of 'team SA behaviors,' which can be numerous depending on the context; these flow into the second phase, wherein various scenarios are introduced via several 'events.' The third phase stratifies 'acceptable responses' across various categories of task-specific behavior. The results of the third phase are fed into scripts of each of the events that occurred within phase II. Finally, within phase five, a structured form is created including information about the scenario, scenario-specific responses, a 'code' (unspecified), and a 'hit.'

Strengths of SALIANT are many. First, the exhaustive, extensive data collection provides a wealth of objective data. Studies with SALIANT indicate it ability to measure:

- Demonstrated awareness of the surrounding environment
- Anticipated need for action
- Demonstrated knowledge of tasks
- Demonstrated awareness of information

Of course, the exhaustive and extensive requirements for successful application of SALIANT also exist as a disadvantage for prospective users who do not have the time, money, or resources. Further, it requires extensive pretest setup as well as the use of trained observers. These concerns suggest against its use in the current experiment.

Real-time Probes. Developed in the United Kingdom (Durso, Hackworth, Truitt, Crutchfield, Nikolic, and Manning, 1998), real-time probes query operators for knowledge of specific aspects of a given situation at all levels of SA, similar to SAGAT, but provide probes one at a time during ongoing operations rather than during simulation freezes. That is, testing is stopped at random times to yield a voice rating. Testing is continuous and, as all information is available for operators to refer to, time to respond is used as the measure of SA (unlike freeze techniques, which check for the accuracy of the answer). Real-time probes are similar to SAGAT in that the probes utilized result from an SA requirements analysis (Jones and Endsley, 2000).

Real-time probes are strong measures because they retain objectivity and are applicable to any complex system. Further, they are related to be timely, simple, and easy to administer and use (Gawron, 2002). Limitations in their use include the need for interruption of real-time, human-in-the-loop testing (discussed below), they require appropriate queries, and they are not yet validated.

Selection of an appropriate SA measure for the current research. Endsley (1998) found the SART scale to be highly correlated (R^2 = 0.67 to 0.74) with the simple subjective SA rating, the evaluation of the sufficiency of one's SA, and the subjective rating of confidence level. Whatever subjective impression was being tapped by these scales, they appeared to draw upon much the same factor, according to Endsley (1998). Direct analysis comparing SART and SAGAT scores revealed several items. Component and correlation analyses suggested that the 13 SAGAT variables were independent, meaning that there was no support for trying to compile SA queries on different

situational aspects into one combined SA variable. As a result, each SAGAT variable was treated independently in comparing the SART score. Regression of SAGAT variables on SART was found not to be significant on any component, and there was no relation between subjective SART ratings and any of the SAGAT variables. SART component examination also found no correlation with SAGAT measures. Subjective assessments of SA derived from SART were not related to the objective measures of SA provided by SAGAT (Endsley, 1998).

The aforementioned study supports the utility of using a test-battery approach for the evaluation of display concepts (the independent variables of the study in question, the same as would be in SATS-like investigations). SART provided information as to *why* the performance results were as they were, not simply that 'a' was more efficient than 'b'. The finding is related as useful in that SART could be used in actual flight operations in the evaluation of design concepts when detailed performance measures are unavailable.

SAGAT was found to provide further diagnosticity with respect to changes in SA resulting from the display concepts in the study. Endsley (1998) posits that this finding most likely reflects changes in attention allocation and processing with the new displays. In short, SAGAT was able to reveal some hidden tradeoffs that arguably could not have been revealed with performance measures alone. That is, SAGAT reveals SA effects that may prove important, at least for the complex mission performance in the study for which it was evaluated. However, Endsley (1998) cautions that, since SAGAT scoring is based on binomial data (correct or incorrect), more data is needed to reach a level of statistical significance than might be required with other measures.

It is interesting that there was no correlation between SART and SAGAT. However, the fact that SART, a subjective measure of SA, was highly correlated with confidence level in SA and subjective performance has interesting implications for the measurement of SA. The fact remains that, since subjective and objective measures of SA showed no correlation, doubt is cast on the validity of subjective SA measurements as indicators of an operator's actual SA (Endsley, 1998). Endsley (1998) further states that the SART ratings found were skewed to be more reflective of increases in SA on some factors (e.g., those items for which the SAGAT scores were higher) as opposed to others. If this is indeed the case, it suggests that subjects may not be aware of other changes in their own SA that may be induced by a particular design, and that evaluations of such data should be conducted with caution. Endsley (1998) concludes by stating that, since the SART scores were found to be so highly correlated with confidence level and subjective performance, subjective SA ratings should be viewed as good indices of these aspects, but perhaps not 'veridical' (true) representations of SA. Similar results were found in studies of ATC (Endsley, Sollenberger, and Stein, 2000). However, in at least one study (Jones and Endsley, 2000), response accuracy as well as response time was measured in order to compare the real-time probe with both SAGAT and SART in an effort to gauge its sensitivity. The real-time probes were found to be insensitive when taken individually, but, when combined to create an 'overall indicator', the measure showed significant sensitivity. As a result, the researchers suggest that the number of repeated measures be increased (increase the number of times each individual question type is posed) during a scenario might increase sensitivity. The SART was again found (as in the study above) to be significant only with this 'overall indicator'; SAGAT's

individual queries were again found to be independent (i.e., aggregate scores are not useful, but did provide diagnosticity)(Jones and Endsley, 2000).

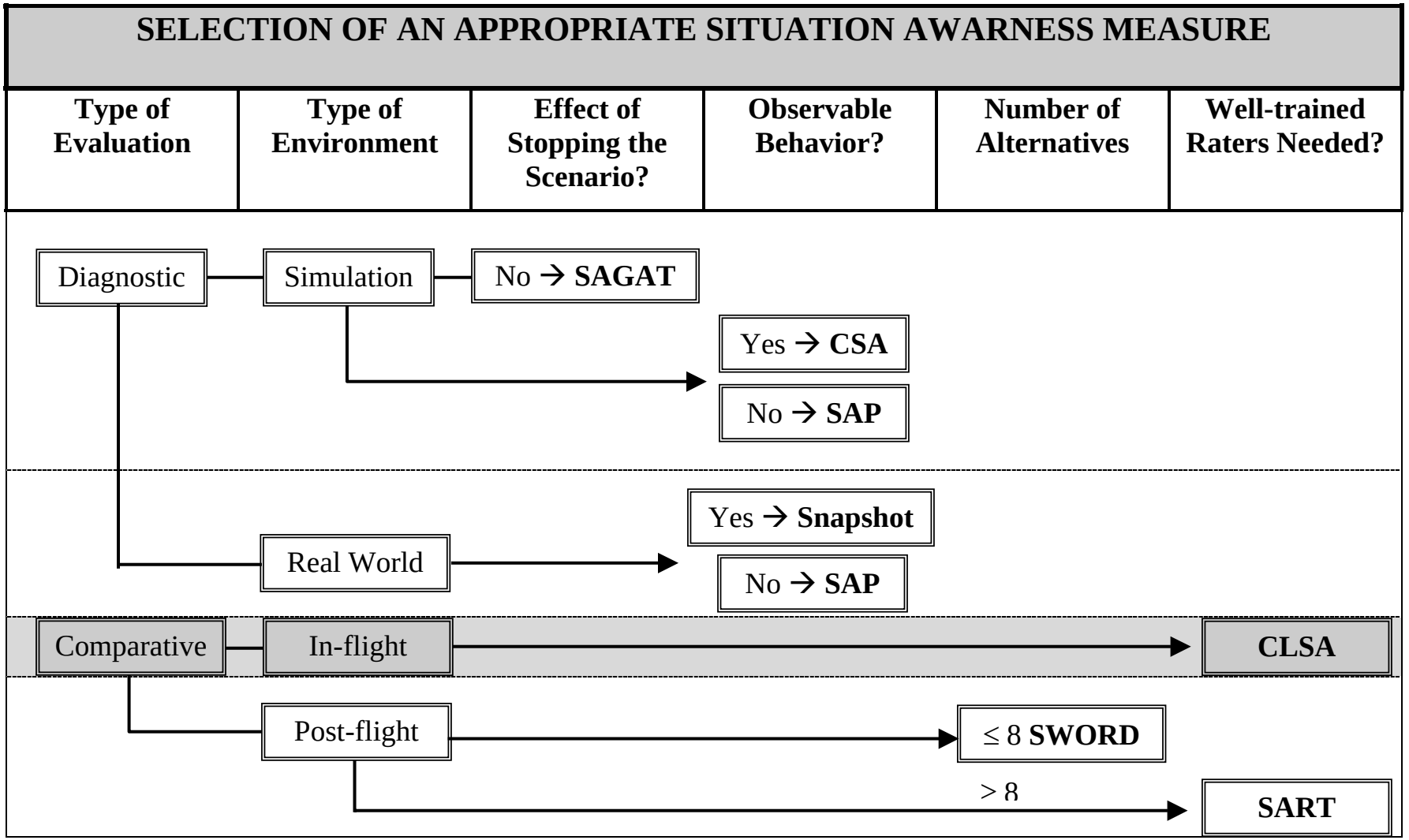
A weak but definite correlation was found between the real-time probes and SAGAT, unlike SART probes (Jones and Endsley, 2000). These findings suggest that, at least at some level, real-time probes are measuring SA. Additionally, both accuracy and response time for real-time probes were shown to be of value. The researchers also chose to compare both real-time probes and SART with a subjective workload metric, the NASA Task Load Index (NASA-TLX). A weak correlation was found, suggesting that further research is warranted. The SART/NASA-TLX result isn't too surprising, since SART contains a strong workload component (supply and demand of attention), but the real-time probes do not. Subjective results indicated that real-time probes weren't intrusive, and, given the sensitivity results, this suggests that events can be increased to bolster sensitivity without an intrusive effect.

Shortcomings of all techniques, as related in the literature (Metalis, 1993) include the indirect nature in which they attempt to measure an individual's SA. They are indirect because there are so many other variables, such as a pilot's quality of training, talent, and the operating vehicle itself (be it a simulator or real aircraft). Metalis (1993), much like in the investigations of Endsley and others discussed above, attempted to overcome these limitations by using a statistically weighted combination of measures in an effort to improve sensitivity with similar results.

Table 11 provides an overview of all salient SA measurement techniques and, considering the nature of the current research with respect to the limitations described above, it appears that both SAGAT and SART are candidates for inclusion. Of course, numbers that are more objective can be derived from performance aspects of the flying task, but relating these to SA requires an inferential leap. For example, Andre, Wickens, Moorman and Boschelli (1991) found that the time and the effectiveness of recovery from an unusual attitude was a useful measure to differentiate between candidate avionics displays. Perhaps a similar extension can be applied to future GA display investigations.

As real-time probes are designed for use in situations wherein freezes are not feasible (as in real flight) and since they show some correlation, one wonders whether they can be used in simulations (particularly future GA simulations) since the SART had no correlation to report at all with the objective SA measure (SAGAT). This is an ongoing research question. It seems the literature does not as yet support the use of a combined SAGAT/real-time probe approach, especially within simulation studies. However, it appears that a need exists to investigate its use and provide further evidence as to its usefulness (or lack thereof).

TABLE 11. Selection of an appropriate SA measure for the current experiment (From Gawron 2002, p. 15-2)



Workload

Workload defined. The term 'workload' has been used to describe elements of interactions that occur between an operator and assigned tasks (Gopher and Donchin, 1986). Further, workload is described as a measure of the *cost* of accomplishing those assigned tasks for the human; costs include fatigue, stress, and the depletion of attentional and cognitive resources resulting in the inability to accomplish additional tasks and often concomitant performance decrements. Put another way, workload is related to differences between resources that are available and resources that are required by a particular task or situation (Sanders and McCormick, 1993). Thus, workload is inversely related to reserve capacity; similarly, it is also directly related to the required level of task performance (Wickens and Hollands, 2000). Workload drives requirements (as in systems engineering), which are met with human capabilities. As humans by nature are such variable creatures, it follows that the human/operator strategy cannot be 'straightforward'; that is, the variability that exists amongst individuals in their ability to meet cognitive and physical demands makes evaluations of workload necessary and imperative in a system, and one must try to account for and support these variations. As such, designers and evaluators come to the realization that performance is not a panacea in system evaluation—it is just as important to adequately consider the task demands that are imposed on an operator's limited resource pool.

Research in mental workload can be said to explore three areas: 1) prediction of performance, 2) the assessment of workload imposed by the system on the operator, and 3) the assessment of workload that is experienced by the human. Such investigations can

provide empirical evidence for function allocation and for the monitoring of operators for task adaptation. Additionally, they can also provide a means to select operators who possess higher mental workload capacities for demanding tasks, and can provide for equipment and technology comparisons with respect to imposed workload, which is a focus of the current study (Wickens and Hollands, 2000).

Workload theory. The main concern in dealing with workload is that there is a lack of a single general theory with which to guide its measurement (Meister, 1995). As such, there are many theories of mental workload that can be found in the literature. These theories generally relate an assumption that the human operator can only be engaged in one task at any one time, and, when a multitasking situation presents itself, the human must prioritize and choose which task to perform first and decide when to change over to a different task. This has been postulated as a result of the limited capacity (or limited-channel) model (Wickens and Hollands, 2000; Salvendy 1997; Kantowitz and Knight, 1976a, 1976b, 1977b). That is, there exists a limited resource pool from which to draw in the performance of tasks. In mental workload studies, tasks are usually labeled as primary, secondary, physiological, and subjective estimates (Gawron, 2002). Primary and secondary measures are discussed first followed by a discussion of physiological measures and subjective estimates.

Primary tasks. Primary tasks, such as the performance resource function (PRF) overall system measure, are simply measures of task performance. The general assumption with respect to primary measures is that, as workload increases, the additional processing requirements required by that workload increase will lead to degradation in

performance, but this has been shown to not always be true. However, concerns with primary task measures are: a) that the primary task of interest may require relatively few resources, resulting in perfect performance (i.e., a 'data-limited' condition), b) that two different primary tasks may be substantially different in how they are measured and what those measures mean, and c) that it may be difficult or impossible to obtain measures of performance using the primary task. Primary task measures are typically used in conjunction with criteria that specify a range of acceptable performance and represent the only means of assessing the adequacy of operator performance within a given system (Wierwille and Eggemeier, 1993).

Secondary tasks. Secondary tasks investigate spare capacity; that is, what is left of the resource pool due to its contents not being directed to performance of the primary task. It is those 'leftover' resources that will be used by the secondary task. Put another way, secondary measures provide useful information on the *low end* of the workload continuum, where primary task measures are notoriously insensitive. These measures require subjects to perform the primary task, within that task's specified requirements, and to use any spare attention or capacity to perform a secondary task. Any decrement in performance of the secondary task is operationally defined as a measure of workload (Gawron, 2002). Advantages of secondary task measures include the fact that face validity is high (i.e., the measure *appears* applicable and relevant) and thus has some diagnostic ability (a concept discussed in detail later), and that a single secondary task can be applied to two primary tasks. Further, they may provide a sensitive measure of operator capacity and distinguish among alternative system configurations. Secondary measures may thus provide a sensitive index of task impairment due to stress, and may

even provide a common metric with which to compare different tasks. Disadvantages of secondary measures are that these tasks are specific to resource pools (as noted above), and that the pool that is tested may not be appropriate for performance goals. Wierwille and Eggemeier (1993) caution that secondary task intrusion should be carefully evaluated because of the possibility of the operator requirement to modify the allocation of processing resources to the primary task. That is, the structure or design of the primary task may affect the sensitivity of the secondary one; similarly, the secondary task may interfere with the primary one.

To explore this relationship further, consideration is given to the pioneering work of Kantowitz and Knight (1977). In their testing of tapping timesharing studies involving 'easy' and 'difficult' primary tasks, they found that when the primary task is performed by itself (e.g., task 1 only), performance was better for the 'easy' version of the task. The tapping timesharing task involved the movement of a stylus between left and right target plates in time with pacing lights, which flashed from side to side at two flashes/sec. The differences between the 'easy' and 'hard' variants of the task were in the widths of the targets (5.08cm for easy and 1.27cm for hard) and in the movement amplitude of the lights. The implication is that the *same* task can be manipulated with respect to difficulty without changing the qualitative nature of the task. As predicted by the limited-capacity model, a secondary task which also draws upon the resources of the channel has little effect for the 'easy' variant; the demands imposed by both together do not exceed available channel capacity so that little or no decrement is observed in the primary task. Alternatively, if the secondary task is combined with the 'hard' variant of the primary task, then the total processing requirements for both tasks exceed the available channel

capacity, which is revealed as a decrement in primary task performance (Kantowitz and Knight, 1977) (see Figure 5). Capacity increases with momentary task demands, but spare capacity that is available for a secondary task decreases as the primary task demands and receives additional capacity.

Interestingly, Kantowitz and Knight had to re-evaluate their analysis when they obtained dual-task additives (essentially parallel lines in Figure 6 as opposed to the interaction suggested in the graph). This finding was consistent with *stage analysis* (e.g., considering tapping information or digit complexity as used in their study), which clearly rejected both the model of limited capacity and variable-allocation capacity models, but the single- vs. dual-task interaction of Figure 6 was not compatible with a simple stage model. This 'simultaneous finding' of additivity for both dependents ruled out capacity tradeoffs. That, and findings in a subsequent experiment evaluating tradeoffs in vision and audition, forced Kantowitz and Knight (1977) to change their model from limited-capacity to a 'mixed parallel model' with both limited capacity and stage features. It was clear to them that this 'hybrid' model, although more general, could account for the results of both experiments. The rationale is that a source of limited capacity feeds both a *response stage*, which controls the outputs for both component tasks involved, and at least two earlier stages, each associated with one of the component tasks (Kantowitz and Knight, 1977).

Division of capacity between the two earlier stages is 'fixed' by considerations such as instructions or pay-offs establishing the relative importance of primary and secondary tasks. These two early stages cannot tradeoff capacity dynamically. As a result, manipulations of difficulty affecting different stages will not interact.

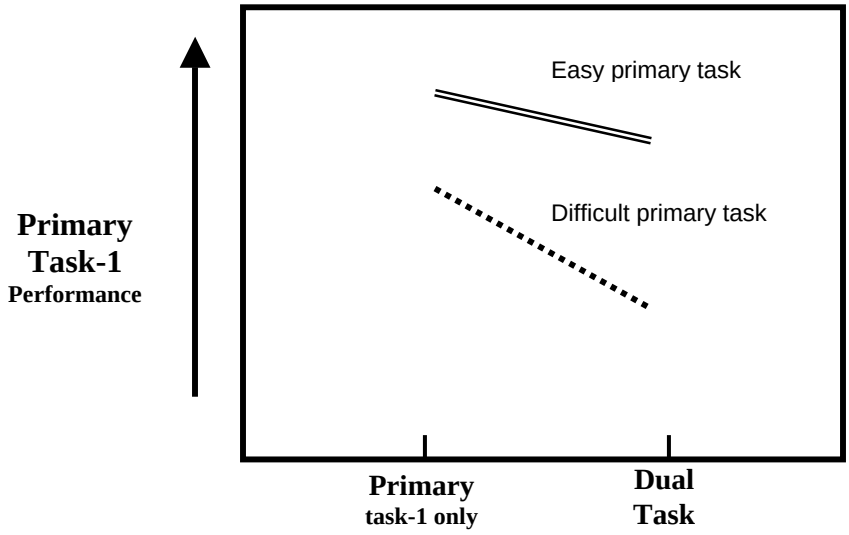


Figure 6. Predictions of limited-capacity model when single- and dual-task conditions are compared. (From Kantowitz and Knight, 1977, p. 345).

In summary, then, when response execution requirements are increased, the output stage is not granted sufficient capacity to process both component tasks without impairment. This causes the model to mimic a limited-capacity system, which results in the traditional interaction seen in Figure 6. However, when only the dual task environment is considered, the hybrid model looks like a stage model because the early stages cannot trade off capacity with each other (Kantowitz and Knight, 1976); see Figure 7.

It is hoped that this introduction to attentional and processing models diagramming resource limitations will provide the basis with which to explain workload measures as relates to future candidate GA system evaluations. Any secondary tasks, if utilized in such evaluations, should of course be germane to piloting (i.e., keying of the microphone or changing plan forms of a global positioning system [GPS] display), and they should occur often enough to suggest a reasonable indication of spare capacity.

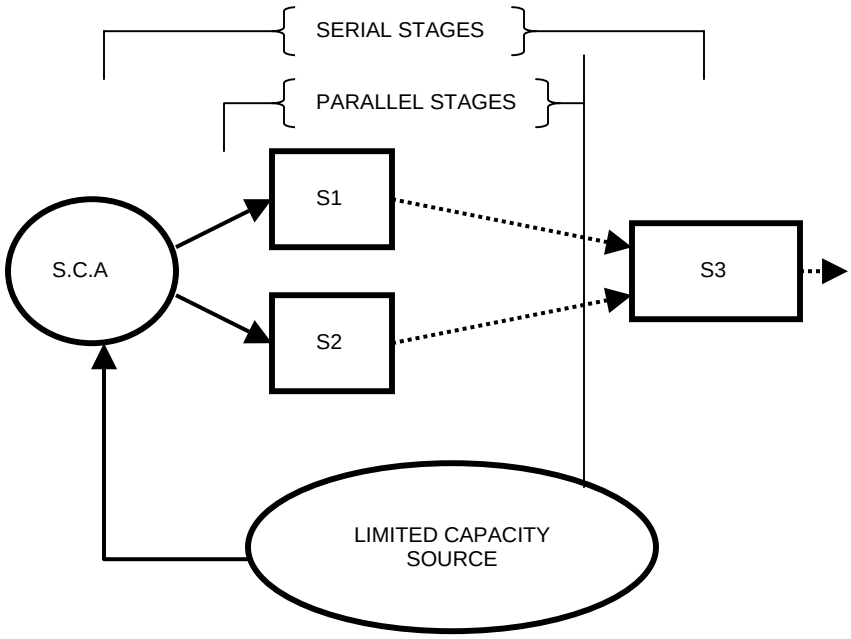


Figure 7. Schematic rendering of a hybrid model. (From Kantowitz and Knight, 1977, p. 359).

Two of the four main categories of workload measures have already been mentioned (i.e., primary and secondary), but two others exist: physiological measures and subjective measures. These latter two will be discussed at this point.

Physiological measures. Physiological measures (sometimes called *psychophysiological assessment methods*) offer an attractive route to circumvent other obtrusive measures. These are defined as "manifestations of workload or increased resource mobilization through appropriately chosen physiological measures of autonomic or central nervous system (CNS) activity" (Wickens and Hollands, 2000, p. 465). Most physiological measures of mental workload are predicated on a single-resource model of information processing as opposed to a multiple-resource model (Sanders and McCormick, 1993). This means that information processing involves activities of the CNS, and this activity can be measured. Another way to consider this is as task demands

change and the operator adjusts the level of mental activity germane to that task's performance, there are associated changes in the operator's physiological systems (Salvendy, 1997). Several physiological measures have been investigated in the literature; however, it is thought that more involved discussions of particular research initiatives with respect to aviation physiological workload measures are better suited for the current research and are thus discussed in subsequent sections. Suffice to say, physiological measures offer distinct advantages. As mentioned, they do not (ideally) affect the task at all: what they do is provide objective information that is quantified in physical units. Further, they usually require no additional activity or performance from the subject. Disadvantages using these measures are that they are generally insensitive, are highly variable, and are subject-specific. These concerns will be presented as relates to aviation shortly.

Subjective measures. Subjective measures of workload have been described as "coming closest to tapping the essence of the concept" (Sanders and McCormick, 1993, p. 82). These measures come in the form of rating scales. Ratings consist of an ordered sequence of response categories that define the correspondence between the stimuli and the responses. It should be noted that there are no direct relationships between values on any workload scale and specific measurable phenomena. Thus, most scales provide ordinal data that is used for indexing the relative differences perceived rather than absolute levels. However, and as Salvendy notes (1997), a particular rating scale either asks for a single *unidimensional* rating concerning the overall workload level or ratings on multiple dimensions for each task condition, which are easier to obtain, or multidimensional ratings are used, which can provide diagnostic information that

spotlights the nature of the workload. These ratings can be either absolute or relative, wherein operators are asked either to compare the condition of interest to a single standard or to multiple conditions of interest (also known as redundant). An example of a unidimensional rating scale is the Modified Cooper-Harper (MCH) scale (Wierwille and Casali, 1983). Others have taken the view that subjective workload, much like the resource theory concept itself, has several dimensions; that is, it is *multidimensional* (Wickens and Hollands, 2000). Examples of multidimensional scales are the NASA-TLX, which assesses workload on each of five 7-point scales, and the subjective workload assessment technique (SWAT), which measures workload on three 3-point scales. A third type of rating scale is *hierarchical*, which separates the evaluation process into a series of explicit decisions. Hierarchical scales do provide selectivity, but they do not provide diagnosticity.

Regardless of which scale is used, their use is advantageous in that they provide an integrated summary from the operator's perspective, which can arguably be done through no other avenue. Additionally, subjective measures are the most direct method for evaluating the *human cost* of task performance – the relation between the demands of a task and the resources available to support it. Some disadvantages in using rating scales are that the ratings are limited to observable actions, task requirements, system performance, and environmental factors. Further, use of a scale cannot allow inference to stress or other psychological consequences of performing a task. All three scales will be discussed in some detail as relates to aviation.

Criteria and categories for workload measures. There exist many issues with respect to workload measurement. Meister (1995) relates that there is an inverse

relationship between measurement control and operational realism. Additionally, the fact that behavior is multidimensional needs to be considered when choosing a particular measure, and even then, the relationship between objective and subjective data is unclear. Designs that attempt to address external validity must still grapple with the difficulty of generalizing any results to the real world. Cognitive tasks are difficult to measure, and the determination of the contribution(s) of each component being measured to the overall system performance is likewise a tough task.

Nevertheless, any discussion of workload would remiss if it did not delve into the indices that impart specific meaning and description. Useful measures of mental workload should meet the following criteria (properties):

- *Sensitivity*: A criteria referring to how well a measure detects changes in the mental workload (e.g., task difficulty and/or resource demand). The measure should distinguish task situations that intuitively seem to require differing levels of mental workload (Sanders and McCormick, 1993). Oftentimes, the degree of sensitivity of a given measure depends on that workload level experienced by the operator. It is thought that performance measures are insensitive at very low levels of workload where adequate spare capacity exists to meet task demands—even if they increase. Performance measures are more sensitive at higher levels of workload where adequate spare capacity exists to meet task demands—even if they increase. Performance measures are more sensitive at higher levels of workload wherein the limits of the operator are quickly being reached and,

consequently, performance deficits are expected to occur with a further increase in task demand (Salvendy, 1997).

- *Diagnosticity*: This criterion indicates not only when workload varies but also the *cause* of that variation. When considering multiple resource theory, the measure should indicate which of the capacities (resources) are varied by demand changes of the system; such information allows the implementation of better solutions as a result (Wickens and Hollands, 2000). For example, by varying only one aspect of the task at a time, elucidation of which particular aspect is responsible for the non-optimal workload level may be fostered (Salvendy, 1997).
- *Selectivity*: When a measure is sensitive to *only* changes in capacity demand, it is termed 'selective.' Another way to consider this criterion is it has 'no noise.' That is, the measure should not be affected by items generally considered not to be a part of workload (e.g., emotional stress or physical load)(Sanders and McCormick, 1993).
- *Reliability*: The measure should offer a reasonable estimate of workload with a 'bandwidth' such that any important, transient changes can be observed. The reliability criterion is concerned with whether the measure is *stable and consistent* over an extendable time. This is a question of being able to replicate workload measures in similar environs: if they can be replicated, they are considered reliable. As such, those measures that fluctuate wildly, independent of task dimensions and workload, have no predictive value and are therefore not useful in any meaningful way (Salvendy, 1997). There are many ways to test the repeatability of a particular measure, such as *test-retest reliability*, use of

alternate forms and *split-half* methods, and evaluating *internal consistency* (such as inter-rater reliability with the use of expert raters)(Gawron, 2002).

- *Obtrusiveness*: Sometimes called 'acceptability', this criterion indicates a measure that does not interfere with, contaminate, or disrupt performance of the primary task. In short, the measure should be acceptable to the person being measured (Sanders and McCormick, 1993). Salvendy (1997) relates it is always a useful practice to fully explain the nature of the measures as well as answer any questions that operators may have about not only the procedures but also later use of the data. The collection of data should not attract the attention of the subject, which may affect that subject's workload. The operator must *accept* the measurement procedure. Indeed, if the procedure asked of participants is unpalatable to them (in whatever manner), they will likely not provide the researcher with good, solid data.
- *Validity*: There are many types of validity with respect to workload, but a general definition of validity is that the measure measures what it is supposed to measure. A measurement technique is considered 'valid' whenever it can discriminate between *easy* and *difficult* sessions of the scenario (Veltman and Gaillard, 1993). *Face validity* is assured when, for example, one receives SME confirmation of a particular environment or testing regimen. Additionally, consideration must be given to the face validity from the viewpoint of participants; that is, does the environment or testing regimen *look* as if it is germane and applicable 'on the face?' *Concurrent validity* is assured through an adequate correlation (i.e., $r \sim$ 0.60 or higher) between two more measures of workload. *Content validity* is

concerned with the desirable quality of measuring all of the important aspects of a particular situation, and no irrelevant ones. This type of validity can be difficult to design into an experiment and typically requires cooperation and understanding amongst involved researchers to ensure it. Perhaps this can be assured, in the current context, through the support and oversight of a dissertation committee! *Construct validity* is difficult to measure depending on what construct or model the investigator is seeking to apply the results to. Assurance that, for example, pilots exert more effort as workload increases is good construct validity. Finally, *predictive validity* ensures that measures that are taken in one environment predict those taken in another environment.

There are other considerations as well within the arena of workload measurement. Data must be collected in an unbiased fashion, in either sampling or representation, for reliable data to result. For example, when considering VFR-rated pilots involved in a display design or evaluation experiment, a researcher should ensure that the sample utilized represents both novice and expert pilots; that is, one must sample adequately from within the continuum of experience that indeed exists in the real GA world. Accuracy and precision are supported when a particular measure results in data that is both correct, *and* that it repeats that way. One must be able to measure specific values or ranges of values such that they are correct, repeatedly. Ideally, measures should strive for simplicity; simple rather than complex measures should be utilized, and the timeliness of the task must be sufficient at completion.

Mental workload measurement in aviation systems. The task of flying an aircraft involves the time-sharing of several tasks; i.e., it is a multiple-task situation that places a great deal of mental workload demand on the pilot. Mental workload of pilots is often high due to the complexity of the flying task. Besides flying an airplane, the pilot has to navigate, communicate and monitor the system (Veltman and Gaillard, 1993). Historically, these tasks have been collectively phrased in their ideal order of operation; that is (as mentioned), 'aviate, navigate, and communicate.'

Mental workload is an elusive concept that cannot be tackled by only one measurement technique (Veltman and Gaillard, 1993). This is especially true for measures occurring within flight operations. The problem with workload research is that there are no other criteria against which measurement techniques can be validated. The only way to validate techniques is to minimize the influence of other factors that determine mental workload. Additional considerations in flight environs are that of age and expertise. A frequent finding is that expertise improves performance of domain-relevant tasks by reducing workload demands on short-term memory (STM) capacity (Lassiter, Morrow, Hinson, Miller and Hambrick, 1996). This result can be explained by consideration of the background of the pilot; specifically, the pilot's ability to draw on generalized background knowledge and ability. To access the store of knowledge in the performance of a domain-relevant task, less effort may be required by the expert when compared to the novice, who may not yet possess the requisite knowledge in a form that is as useful for application to the required task. The researchers suggest this finding may result because the knowledge possessed by the novice is far less structured and less 'automatized' than that of the expert (Lassiter et al., 1996). This finding may have an

opposite effect on research investigations into future NAS efforts such as the SATS in that prior knowledge by experts may actually hinder performance because the system has operational aspects that differ so much from current practices. At least in one investigation (Lancaster et al., 2003), as mentioned, it was suggested that experienced pilots presented superior performance over novices with respect to glide slope (GS) manipulations that are envisioned in the SATS effort.

Workload has also been evaluated, albeit sparingly, within commercial operations that utilize data link. When evaluating data link text and digitized speech presentation types, Rehmann (1996) found no significant differences in subjective effort ratings between data link conditions, indicating that the method of ATC communications did not affect crew workload in any measurable way. Conversely, within the digitized speech conditions, the PF glanced at the data link unit significantly less often and glance duration was shorter, suggesting that the aural annunciation of ATC data may indeed decrease workload. As data link is sure to be a component of future aviation operations, data with respect to presentation of ATC information within these systems and their affect on workload appears to be sorely needed, especially within GA operations involving a single pilot. As mentioned, most locatable data link research has focused solely on commercial operations involving flight teams, further spotlighting this need.

Subjective workload measures in aviation systems. Subjective techniques are almost always sensitive to changes in task load in both the simulator as well as during real flight. This indicates that subjective techniques are reflecting changes in task load rather than changes in effort. Thus, subjective and physiological techniques do not give the same kind of information and are therefore both important for mental workload

studies (Veltman and Gailliard, 1993). In general, subjective techniques appear appropriate for a wide range of test and evaluation research investigations, from momentary to long-term, and are expected to yield high levels of sensitivity, and, in appropriate applications, some indication of diagnosticity (Wierwille and Eggemeier, 1993). However, claims have been made that the reliability and validity of subjective ratings of mental workload are insufficient. The reasoning behind this assertion is that all mental processes are not 'introspectively available', and accordingly, the subjective measures can possibly yield an underestimation of workload (Svensson, Angelborg-Thanderz, and Sjoberg, 1993). One must also take into consideration the fact that, since subjective measures by definition depend on subjective judgment and reporting, they can therefore be influenced by factors other than the actual levels of loading experienced by the operator (e.g., new system biases or context effects). The latter can be overcome by ensuring that all operators have similar relevant experience and perform under an equivalent range of system conditions; the former can be very difficult to address (Wierwille and Eggemeier, 1993). Many evaluations of various rating scales have indicated that, especially with respect to sensitivity and operator workload measures, the Modified Cooper/Harper, SWAT and the NASA-TLX are powerful tools indeed (Wierwille and Eggemeier, 1993). Each of these scales will be discussed in turn.

The NASA-TLX is a multidimensional scale consisting of six dimensions: mental demand, physical demand, performance, effort, and frustration. The subscales have to be weighted by means of a pair-wise comparison of the subscales. These weights are used to calculate the overall workload score. It has been proven a sensitive measure to indicate workload in several studies (Veltman and Gailliard, 1993). Bauschat (2001)

relates that the NASA-TLX had good acceptance by evaluation pilots in his study. The TLX scale has been successfully demonstrated with respect to sensitivity in several flight experiments in which demand manipulations were incorporated (Wierwille and Eggemeier, 1993). Additionally, the TLX scale has been noted as particularly useful for applied applications and has been considered potentially more sensitive at low workload levels than is the SWAT (described below). A downside to its use, as noted in Wierwille et al. (1993), is that it requires data be gathered from operators, which may be time-consuming. Still, it has been reported that the TLX enjoys higher user acceptance than does either the SWAT or the MCH scales (Wierwille and Eggemeier, 1993).

An avenue for the determination of the handling qualities of an aircraft has typically been based on the Cooper/Harper rating scale developed by Cooper and Harper in 1969. A pilot can give a handling quality between 1 (excellent) and 10 (severe deficiencies). Wierwille and Casali (1983) discovered a way to use the Cooper/Harper scale to determine mental workload by modifying it to become the *Modified* Cooper/Harper (MCH) rating scale. Specifically, they retained the original scale, but changed the verbal descriptors in an effort to facilitate use in a wider variety of workload applications, including those with a cognitive aspect (Wierwille, Rahimi, and Casali, 1985). In their scale, '1' stands for an easily solvable instructed task (operators' mental effort is minimal and desired performance is easily attainable). Accordingly, '10' means impossible (the instructed task cannot be accomplished reliably). See Figure 8 for a graphical depiction of this scale. Bauschat (2001) relates that the MCH had good acceptance by evaluation pilots in his study. Wierwille et al. (1985) presented a paper in which they evaluated 16 mental workload assessment techniques utilizing a moving-base

flight simulator in an effort to gauge the relative sensitivity and intrusiveness of those measures.

Results indicated significant sensitivity of the MCH scale with respect to mediational loading. Specifically, the MCH was found to be linear; that is, the scale can discriminate between the three load levels utilized in the study quite well, even with very few subjects (Wierwille et al., 1985). Further, the MCH has been successfully applied to workload assessments in several flight experiments subsequent to its inception that utilized several types of demand manipulations (Wierwille and Eggemeier, 1993). What makes the MCH particularly useful in aviation research is that it results in no differential intrusion (thus supporting non-obtrusiveness) relative to other performance-based and physiological assessment procedures that have been evaluated (Wierwille and Eggemeier, 1993). Further, and as Wierwille and Casali (1983) contend when they studied various metrics for the measurement of workload specific to flight tasks, the MCH was found to be particularly sensitive with respect to changes in *communication load*. This finding suggests that the MCH elicitation technique could be usefully applied in studies involving airborne data link and variations of data link and their effect on perceived workload.

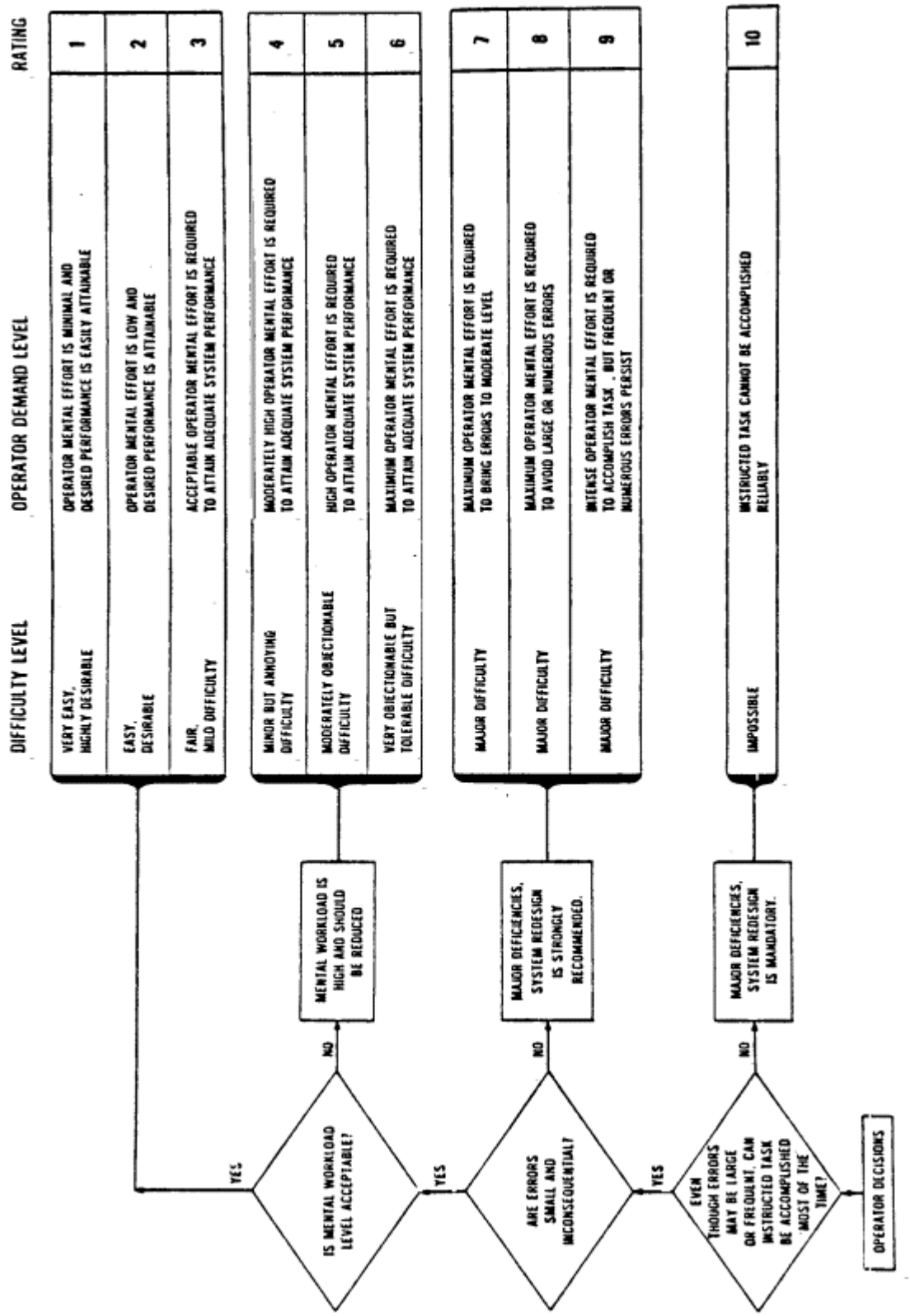


Figure 8. The Modified Cooper-Harper workload rating scale (Adapted from Wierwille and Casali, 1983).

Subjective Workload Assessment Technique (SWAT). The subjective workload assessment technique, another multidimensional scale, has also been extensively employed in the assessment of pilot workload. Its sensitivity has been successfully demonstrated in demand manipulations of flight environments. SWAT has been "viewed as having the greatest potential for identification of factors such as cognitive mechanisms affecting mental workload judgments" (Wierwille and Eggemeier, 1993, p. 267). Another attractive ability of the SWAT is its capability to derive interval scale values (thus permitting parametric statistical analysis) from a validated scaling technique and assign *meaningful* numbers to a pilot's subjective impression of workload (Hale and Piccione, 1992). The SWAT, however, has a downside in that it requires substantial data to be gathered from the operator; indeed, Wierwille et al. (1993) relate that the scale development procedure can require up to an hour to complete, and this may affect user acceptance. Furthermore, SWAT lacks a threshold criterion value above which workload can be said to be 'unacceptable.' Without such a definitive 'critical' SWAT value, the most applicable use of SWAT, according to Hale et al. (1992), may be in comparing alternative system designs rather than attempting to define unacceptable levels of workload associated with the employment of a single system. Nevertheless, SWAT has proven feasible for utilization during operational flight segments, wherein the operator verbalizes ratings at the end of each segment, and these ratings have been shown to correlate well with task analytic estimates (Hale et al., 1992).

Veltman and Gailliard (1993) utilized the BSMI scale (a Dutch workload scale) in their research. The BSMI is a unidimensional rating scale that asks for the amount of effort that has been invested during task performance. Subjects have to respond by

putting a marker on a vertical axis that ranges from 0 to 150. On the right side of the scale are statements like “not at all effortful”, “a little effortful”, “very effortful”, etc. The BSMI results indicated higher scores in their ‘more intensive’ condition of flying curves while performing a continuous memory test (CMT); scores were lower when flying straight and performing the CMT (Veltman and Gailliard 1993). The BSMI is not often used but appears to be a valid and sensitive measure.

Physiological workload measures in aviation systems. Several different theoretical and experimental studies have demonstrated the relevance of physiological measures in assessing mental workload. It has been suggested in the literature that physiological processes represent the interaction between the individual and the environment, particularly task related aspects, and that these processes can collectively be termed *organic cost* (Hancock, Meshkati, and Robertson, 1985). The researchers provide a rationale that underlies physiological measures, based on comments made by Wierwille (1979, p. 577):

> As operator workload changes, involuntary variation occurs in human physiological processes; in consequence, workload may be assessed through the monitoring of the appropriate physiological system. As mental workload presumably affects the activity of the CNS, measures may variously reflect processes such as demand for increased energy, progressive degradation of the system, or homeostatic action mechanisms designed to restore system equilibrium disturbed by such cognitive task requirements.

The Hancock review (1985, p. 1111) presents a definition of mental load as “a reflection of purposive activity in the CNS on the sentient operator.” Veltman and Gailliard (1993) attempted to measure pilot workload with both subjective and

physiological techniques. The techniques were validated by systematically changing the task demands to examine whether the scores received followed these changes. Physiological measures can additionally be used as a supplement to subjective measures, especially when unobtrusive, continuous measures are desirable (Wierwille and Eggemeier, 1993). Types, uses, and experimental results of various physiological indices in aviation workload are described below.

Simple heart rate measures have been said to provide an overall index of general arousal or physical work associated with variations in task demands (Wierwille and Eggemeier, 1993). Measures of Heart Rate (HR) and Heart Rate Variability (HRV) can generally be considered non-intrusive. Indeed, many portable recording devices exist and have been successfully employed in several research investigations. Veltman and Gailliard (1993) relate that the mid-band region (i.e., 0.07-0.14 Hz) has been found to be the most sensitive to changes in mental effort. The spectral energy in this region decreases when the mental effort increases. This region was found to differ significantly from baseline, and the values from the easy sections differed significantly from the harder sections. The researchers (1993) also relate that Heart Rate Variability (HRV) may provide more information concerning mental workload in practical situations where physical activity may play a role because of its apparent sensitivity decrement when compared to physical demands. The fluctuations in HRV depend on the length of the time-window for frequency analysis. A 'long' window results in more stable HRV values but the analysis becomes less sensitive to fast changes in effort.

Wilson and Badeau (1992) relate a note of caution when trying to extrapolate cardiac data from laboratory to the actual flight environment. In their study examining

psychophysiological measures of cognitive workload in laboratory and in flight, they found four to ten percent increases in HR for pilots when performing a laboratory-tracking task compared to a resting baseline. During flight, pilots' HR was found to increase up to 45%. This large discrepancy in percent change suggests that the cardiac system dynamics may well be quite different in these two conditions and may follow different functions. Additionally, Bauschat (2001) maintains that the more demanding a piloting task is, the more significant the difference between the ground-based and real-time and flight-test. Conversely, there have been demonstrations of HRV's ability to discriminate differences in loads imposed by flights tasks rather well (Wierwille and Eggemeier, 1993).

Hancock, Meshkati, and Robertson (1985) suggest that measures pertaining to heart rate and its derivatives are the most practical method with which to assess imposed mental workload. However, since HRV is a particularly sensitive physiological function, it can be vulnerable to contamination from the influences of both the ambient environment as stress. Further, the Hancock group presented a diagram (see Figure 9) depicting the relative strengths and weakness of various physiological measures with respect to 'spatial and systemic congruence' (defined as the actual spatial distance from the CNS of the suggested site of experimental activity) on one axis and 'practicality' (ease of use and cost, specifically in the aviation research arena) on the other. Many physiological measures indicated on the plot were not covered here; the interested reader should consult the Hancock et al. (1985) paper for more information.

Veltman and Gailliard (1993) studied several respiratory parameters:

- *Respiration rate (RR)(min^{-1}):* the number of respiratory cycles per minute;
- *Tidal volume (VT)(ml):* the amplitude of the respiratory signal;

- *Inspiratory flow (IF)(ml/s):* VT/(inspiratory time);
- *Duty Cycle time (DCT):* (inspiratory time)/(total cycle time);
- *Minute volume (MV)(ml):* RR x VT

Their results indicated that all respiratory values differed significantly from baseline values, suggesting increased workload; but the DCT appeared to be the most sensitive. It should be noted here that respiration might also affect HRV. Respiratory frequency and the variability in heart rate around the respiratory frequency are correlated. Veltman and Gailliard (1993) attempted to account for this by calculating a *coherence function*. They relate that coherence is comparable to the squared correlation (explained variance) in a linear regression equation in the time domain.

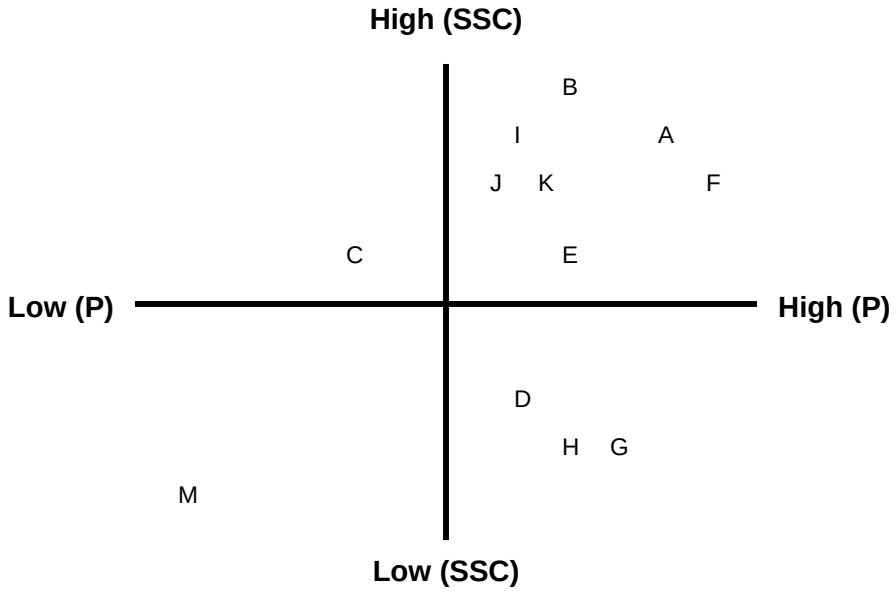


Figure 9. Major physiological measures of mental workload located in two-dimensional space, where (P) = practicality and (SSC) = spatial and systemic congruence. Individual measures are represented by letters as follows: A = ACT, B = Event-related potentials, C = Flicker fusion frequency, D = GSR, E = ECG, F = HRV, G = EMG, H = Muscle tension, I = EEG, J = Eye movement, K = Pupillary dilation, L = Fixation analysis, and M = Body Fluid Analysis. (Adapted from Hancock et al., 1985, p. 1111).

A high coherence means that two signals have a high resemblance for a particular frequency. The coherence in the mid-band was found to be rather low (0.3) compared to the high-band (0.7) and was not found to be different for the experimental conditions. Thus, observed changes in HRV due to task load are not directly related to changes in respiratory frequency.

Electroencephalogram (EEG) involves the recording of evoked potentials from surface electrodes applied to the scalp. Evoked potentials are the small changes in the electrical activity of the brain that are associated with processing of information contained in discrete stimuli. The power of these potentials is quantified within several frequency bands: alpha, beta, theta (which has been linked to SA arousal), and delta. Wilson and Badeau (1992) utilized brain evoked potentials to study the changes in brain activity associated with increasing task demands within an aviation environment. The researchers found that the amplitude of the late evoked potential components decrease with increasing task difficulty. One particular component of the evoked activity, the P2 (a potential occurring at 200msec after stimulus onset) was significantly reduced while the pilot was actually flying the aircraft when compared to ground segments (Wilson and Badeau, 1992).

Strengths of EEG use include reports in research investigations suggesting the ability to demonstrate EEG sensitivity to variations in flight demand (Wierwille and Eggemeier, 1993; Wilson and Badeau, 1992). EEGs are relatively unobtrusive, are diagnostic, enjoy high face validity, and are objective. However, care must be taken in interpretation of EEG data, for it is widely known that various artifacts related to eye and body movements can affect output. Additionally, equipment may malfunction in

vibration and high temperature conditions, and data reduction is time and labor intensive (Gawron, 2002).

The eyes have been called the 'windows of the soul'; this may or may not be true, but the eyes certainly do regulate and process visual input. Eye blink activity interrupts the flow of visual information; thus, visually demanding situations should decrease blinking and shorten the duration of the blinks (Wilson and Badeau, 1992). The researchers relate that military pilots presented the shortest blink rate closures when pilots flew in formation, specifically the wing position. It is suggested that this result is related to the higher visual demands associated with the maintenance of ship position relative to the lead. These results can be expected to repeat, although perhaps not as substantially, in GA simulations that entail in-trail station keeping, wherein pilots must maintain self-separation in landing patterns, especially when the monitoring of a multi-function display (MFD) is involved. Several techniques are available to record eye blink activity, include electro-oculographic (EOG) procedures such as corneal reflex, pupil center corneal reflection distance (which can be hampered in a bright setting), and the double purkinje measure (Gawron, 2002). Blink amplitude is related to situation awareness—increased eyelid opening equals increased SA. Glance duration (length of a generalized glance) is inversely related to SA (i.e., increased glance duration equals decreased SA, likely because the operator cannot figure out what is going on). Fixation duration (i.e., length of a 'visual fix' on a specific area) repeats this phenomenon (Marshall, 1996).

Strengths of eye measures include its relative unobtrusiveness, it is diagnostic, enjoys high face validity, and is objective. However, several aviation investigations in which the author was a part and which utilized eye measures resulted in subject

complaints specific to the rather unwieldy head-mounted device and reflector screen that was used, so perhaps an umbrella assertion of relative unobtrusiveness is somewhat misleading. Disadvantages include (as mentioned) bright light, equipment malfunction in vibration and high temperature conditions, the fact that data reduction is time and labor intensive, and the fact that sensation does not always result in perception (Gawron, 2002).

Following the reasoning of autonomic activation, loads imposed on the CNS will initiate increased activity in that structure. Any increased activity will give off heat, and this is the basis for what has been termed the 'most suitable' semi-invasive measuring site for the observation of such changes—the auditory canal; specifically, the Auditory Canal Temperature (ACT) (Hancock et al., 1985). The Hancock et al., research relates that the ACT is somewhat sensitive to environmental variables such as ambient temperature fluctuations; as a result, they recommend that measures of changes in temperature (ΔT) be utilized rather than absolute temperature. Further, this measure would help mitigate the effect of the inherent individual differences in all physiological parameters (Hancock et al., 1985).

An adrenal-cortex hormone that is active in carbohydrate and protein metabolism, cortisol has been suggested to be a measure of mental stress. Cortisol is obtained from saliva and is analyzed using radioimmunoassays (RIA). Veltman and Gailliard (1993) found that cortisol levels differed significantly between the training day and experimental day in their study of physiological workload measures. Further, during the experimental day, the post-task levels did not differ significantly from pre-task levels. Interestingly, Veltman and Gailliard (1993) relate that previous research has found a relation between cortisol level and a 'defensive personality structure', which they say might have

influenced their results, since, whenever a subject is more defensive, he/she will choose an easier strategy. They conclude that subjects with high cortisol levels perform poorly due to a defensive task strategy.

Mood and mental workload measures. There is no simple relation between pilot emotional state and the characteristics of a particular flight. The same flight can be a challenge or a threat depending on surroundings and circumstances. Svensson et al. (1993) maintain that mere expectations may influence reactions and results. A positive expectation of being able to handle a situation is helpful; negative expectations may reduce tolerance of frustration. Further, stress of any kind can make a threat out of a challenge, and one's mood can tip the scale dramatically (Svensson et al., 1993). Mood has been described (Svensson et al., 1993) in terms of three dimensions: a) hedonic tone (i.e., sad vs. happy), b) tension, and c) activity. They suggest the latter two are the 'most interesting' in the context of flight: high activity/low tension *support* mental performance; low activity/high tension can be said to act in the opposing direction with respect to both mental and physical performance. The Svensson et al. (1993) study found that challenge is indeed a factor driving the coping process and, eventually, performance for flight missions. The aspect of challenge, according to the researchers, increases both problem solving and emotion coping, with the former positively affecting performance and the latter negatively. As a result of their research, the experimenters present a rather new and novel concept that is said to be the opposite of mental workload, termed *mental energy*. They use this concept as a basis of measures involving commitment and activation, and define it as "the ability to regulate successful action in the face of obstacles such as fatigue and fear" (Svensson et al., 1993, p. 991).

Selection of an appropriate WL measure for the current experiment. The preceding discussion outlined the theories of mental workload and their considerations, the qualities of good measures, and presented several examples of measures, especially within aviation systems. Attempts to measure workload within the flight environment are not always feasible, due to safety or other concerns. However, such investigations within the laboratory, especially those utilizing simulation, present an attractive alternative. As mentioned, any attempts to implement cockpit design alternatives must have the foundational support of rigorous human-in-the-loop research. The fact that a speech message takes time to be delivered requires respect to what Simpson and Williams (1980) have called the *system response time*. This measure is defined as the time interval starting with the onset of a signal and continuing until the listener has decided upon and initiated his/her first action. As such, system response time includes such elements as sensation (detection), perception, comprehension, storage, retrieval, and decision-making (Simpson and Williams, 1980). Currently, reaction time is the most frequently cited dependent measure in many investigations of auditory phenomena (e.g., auditory warning research). However, as Baldwin et al. (2002, p. 71) point out, "reaction time may prove to provide only an index of the alerting capabilities of auditory displays while failing to allow examination of more complex issues such as verbal intelligibility and the ability to correctly respond to the message content of the verbal warning". This contention is additionally supported in the driving performance research of Liu (2001) that utilized an auditory display. In that study, 'response time' was defined as the elapsed time between the presentation of information (i.e., the beginning of a verbal warning) and pushing the correct button.

In order for a particular speech display to be effective, the operator must not only respond, but also respond *appropriately*. Indeed, the operator is required to muster additional resources toward the comprehension and selection of the appropriate action, thereby possibly compromising resources that may be applied to other critical tasks. Some of the displays utilized in the current research required the reading of text, thus requiring the pilot to look down at the display to glean information (i.e., 'eyes inside'). Previous discussion has described the concept of *head-down time* and its relation to workload; as such, this index will be included in an effort to provide data for workload determination. As the Modified Cooper-Harper has been validated for use in the cockpit, and indeed was modified from the original Cooper-Harper scale to account for changing trends in automation and the cognitive functions required with them, its appears to be a powerful tool with which to evaluate future GA systems that incorporate such automation, and is therefore a further WL measurement tool for the current research.

It is therefore determined, based on the review presented here, that this battery of workload measures be employed in the assessment pilot workload in future GA operations such as that proposed (i.e., MCH administration and the determination of 'head-down time'). Any secondary tasks, as mentioned, should of course be germane to piloting (i.e., keying of the microphone or changing plan forms of GPS display), and they should occur often enough to suggest a reasonable indication of spare capacity, but it is unclear whether such measures a warranted until the results of the initial experiment are evaluated. It should be noted, however, that mental workload assessments in the aviation domain should be measured in concert with other non-workload measures, such as SA indices, since they play such a vital role in evaluating the ability of the pilot to safely

function in the envisioned operating environment. Thus, an arguably 'well-rounded' picture of the experimental task can be created and analyzed. Any concerns as relates to interference with workload measures are entirely justified; however, at least two situation awareness measures have been shown not to interfere with primary and secondary task assessments, and it is they who are suggested for inclusion in the proposed experiment as well (see SA section). Such indexes of both mental workload and situation awareness, which help to elucidate a universal picture of pilot response within the cockpit, will provide powerful tools with which to based recommendations for future GA systems.

PURPOSE OF THE CURRENT EXPERIMENTS

In order to operate safely, general aviation pilots have long dealt with the need to construct and maintain an accurate mental model of their location relative to other aircraft in the airspace. Within the NAS, GA aircraft must operate safely in locations ranging from the remote to the urban, and share the airspace with traffic that ranges from other GA aircraft to large commercial and military aircraft. The airports used by GA pilots also range from single, non-towered airstrip to multiple-runway, tower-controlled airports with hundreds of operations each day. As GA aircraft performance increases and as the amount and complexity of the available information increases, there is a need to ensure that pilots can safely operate their aircraft while remaining in compliance with existing and future aviation procedures. Increasing complexity due to improving technology is not limited solely to the aircraft. Future iterations of the NAS, such as the SATS, will

coordinate technological innovations of aircraft with advances in airspace management and information systems. Utilizing the latest innovations in avionics, communication, and automation to create functional and operational Personal Air Vehicles (PAV), SATS attempts to offer enhanced services to users of the NAS that include Higher Volume Operations, Integrated Fleet Operations, Lower Landing Minimums and increased single-pilot safety and mission reliability. In order to ensure that users can indeed operate safely and efficiently within this operational context, it is imperative that research initiatives addressing the human element of the system be explored; that is, what are the capabilities, expectations, and limitations of the pilots in these scenarios and how do they relate to various elements of the envisioned SATS system?

Data link technology has been suggested as a viable means of automating information transmission to pilots in the future NAS. Data link typically facilitates the transfer of such indices as aircraft identification, location, and heading—usually visually via text-based displays. Investigations of land-based in-vehicle information systems (i.e., in automobiles and trucks) have demonstrated measurable safety benefits through the use of speech technology, but no one has yet examined such technology applied to GA aircraft. While at least one study (Battiste and Johnson, 2002) has looked into the *technology that provides for data link* (e.g., ADS-B: Automatic Dependent Surveillance-Broadcast or TIS-B: Traffic Information Services-Broadcast); specifically, data link presentation from a purely textual standpoint and its effect on performance, little research has taken the next step to investigate the *modality of those data link transmissions* (e.g., visual vs. auditory). Of those that have looked into modality, none have considered synthetic speech, only digitized speech (Rehmann, 1993, 1996; Rehmann, Reynolds, and

Naumeier, 1993). Another concern is that these studies have typically investigated data link usage among flight crews (commercial or military). The basic premise of such endeavors as SATS is to support *single pilot* operations, and no locatable research has looked into the capabilities of the single pilot with respect to data link and manipulations of presentation. Further, system operability and implementation issues with respect to data link (i.e., crew alerting, message formatting, situation awareness, clearance formatting, and mixed-modality communications) have been specifically identified as critical in many key reports, as discussed previously (FAA 1995; Rehmann, Reynolds, and Naumeier, 1993; Airport Transport Association [ATA], 1991). Pilot situation awareness is a major concern for researchers investigating the effects of data link implementation and was measured in the current research using the two methods mentioned within the SA section (i.e., SAGAT and SART). As discussed, the benefits derived from enhanced situation awareness include: improved safety, reduced workload, enhanced pilot performance, an expanded range of operation, and better decision-making (Regal, Rogers, and Boucek, 1988). Also as mentioned, there is concern that data link may actually *decrease* pilots' situation awareness as a result of the potential loss of 'party line' information currently being obtained from voice radio communications. It has been suggested that the availability of TCAS may provide enough information to adequately compensate for the loss of information. The FAA Technical Center identified the need for adequate situation information in the cockpit and has acknowledged it as a critical research issue. Additionally, the need to assess and resolve the effects of data communications on pilot/controller situation awareness was mentioned in the Information Management and Display thrust of the 1995 National Plan for Civil Aviation Human

Factors: An Initiative for Research and Application. The issue of 'party line induced situation awareness' was ranked number 22 of 45 by the 1991 Air Transport Association (ATA) survey, under the issue of 'Party Line Compensation.' Also, the issue of Crew's/Controller Situation Awareness (issue number 10) was ranked 'serious' by the SAE (Rehmann, 1997).

Research into synthesized and/or digitized acoustic systems has not been conducted utilizing the latest technologies, which have matured in recent years. As such, there is benefit to investigating current speech technologies within the cockpit (e.g., for automated aural ATC directives such as local traffic position or to maintain in-trail station keeping). Such benefits may include increased safety through decreased pilot head-down time (i.e., not having to look down at a textual display or fumble with the radio stack), increased situation awareness, construction and maintenance of accurate mental models, and decreased operational workload. As has been discussed, the DECtalk-powered systems have received much attention in the literature and have had favorable performance results, so justification for their use in the cockpit may be supported. However, advances in microprocessing and the lowering costs of storage coupled with the fact that there is no prior research evaluating newer systems, dictates the need to evaluate a newer system against DECtalk for intelligibility, AT&T's Natural Voices TTS system. Analysis of the results comparing these two synthesizers will provide a rationale for the inclusion of the superior-performing speech synthesizer as a stimulus in the pilot performance experiment to be conducted subsequently.

The design and implementation of non-verbal (e.g., warning tones and alerts) auditory displays has received more attention than verbal displays in the scientific

literature of recent decades. The role of the human is changing rapidly within envisioned future NAS operations, especially one that includes the monitoring of a wide array of displays and warnings. As such, speech displays have been suggested as a means of augmenting performance within such an automated environment, and the current research will help to investigate their use. Indeed, the investigation of different speech modalities (i.e., synthesized vs. digitized) has been called for as an avenue for future research (Baldwin et al., 2002; Harvey, Reynolds, Pacley, Koubek, and Rehmann, 2002). As discussed, previous research investigations of data link presentation have evaluated a combination of digitized/textual modality (Rehmann 1993, 1996, 1997; Rehmann, Reynolds, and Naumeier, 1993) within commercial aviation operations. It is for this reason that a condition of the current experiment was a combination of synthesized/textual modality. Even though these investigations evaluated commercial operations and not GA operations, it was felt somewhat redundant to include a digitized/textual combination. In addition, no locatable research has investigated either synthesized or digitized presentation of ATC directives *alone,* or without a textual reference. As auditory stimuli are more ephemeral and 'fleeting,' as it were, there is merit in studying whether or not pilots can retain within short-term memory (STM) the (typically short) messages from ATC, or if pilots require repeating of the message. This is another question that the current research attempted to answer. Given these shortcomings, the current research sought to provide data in support of future single pilot GA operations that incorporate automation.

To recapitulate, the specific research objectives are to:

- Investigate the maturation of TTS engine technology by comparing, within aircraft cockpit engine noise, a 'newer' speech synthesizer with an 'older' one that has demonstrated superior intelligibility in the past.
- Build upon research that has investigated the effects of advanced communication technology (e.g., data link) on human performance.
- Investigate how different modalities of data link effect pilot workload and situation awareness in the single-pilot GA cockpit.
- Provide recommendations for the integration of mixed-modality displays into single-pilot GA cockpits and in similar systems that present high levels of background noise during routine operations and auditory display presentations.

There are several formal hypotheses related to the current research. These are listed below:

H_1: A 'newer' TTS engine (AT&T's Natural Voices) will result in superior intelligibility within aircraft engine noise at each tested S/N over the 'older' DECtalk TTS engine.

H_2: There is a significant difference in pilot performance across data link modality with respect to the time required to access, understand, and execute ATC data link messages. This variability may be due to data link modality, the age or gender of the pilots, their experience, or to a host of other factors.

H_3: There is a significant difference in pilot performance across flight condition with respect to the time required to access, understand, and execute ATC data link

messages. This variability may be due to flight condition, the age or gender of the pilots, their experience, or to a host of other factors.

H_4: There is a significant difference in pilot workload, both objective and subjective, across data link modality. This variability may be due to data link modality, the age or gender of the pilots, their experience, or to a host of other factors.

H_5: There is a significant difference in pilot workload, both objective and subjective, between flight conditions. This variability may be due to flight condition, the age or gender of the pilots, their experience, or to a host of other factors.

H_6: There is a significant difference in pilot situation awareness, both objective and subjective, across data link modality. This variability may be due to data link modality, the age or gender of the pilots, their experience, or to a host of other factors.

H_7: There is a significant difference in pilot situation awareness, both objective and subjective, between flight conditions. This variability may be due to flight conditions, the age or gender of the pilots, their experience, or to a host of other factors.

METHODOLOGY: SPEECH INTELLIGIBILITY (Experiment I)

Experimental Design – Speech Intelligibility

The two speech synthesizers evaluated in this experiment included DECtalk v4.5 'Perfect Paul' (an 'older' TTS engine) and AT&T Natural Voices v1.4 'Mike' (a 'newer' TTS engine). As discussed in the literature review, the DECtalk product represents the 'best' of the synthesis-by-rule TTS engines, and has been shown to be superior with respect to intelligibility in many studies (e.g., Greene et al., 1986; Ricard and Meirs, 1994). The AT&T product uses 'unit selection synthesis', a synthesis technique utilizing categorically classified pre-recorded speech units, and uses half phones as its basic units (as opposed to diphones). Both speech synthesis systems are included within the Fonix iSpeak v3.0 synthesis software suite. These stratifications result in a 3 X 2 within-subjects design (see Figure 10).

Participants

Ten (10) males and females were recruited to participate in the experiment. A 60/40-gender mix resulted (i.e., 6 males, 4 females). Participants were recruited from the local Blacksburg community, Roanoke, and further outlying areas until the requisite number of participants was achieved (through postings, email, listserv, and word-of-mouth). All participants had pure tone hearing levels (HL) of 10 dBHL from 250 Hz to 1 KHz and 20 dBHL above 2 KHz. Thus, all participants met the requirements of the ANSI standard 'method for measuring the intelligibility of speech over communications systems' (ANSI 3.2-1989).

Independent measures. The first independent variable was representative of the kind of speech synthesizer presented to the participants– DECtalk 4.5 and AT&T Natural Voices 1.4. Each TTS engine was tested at three speech-to-noise (S/N) ratios: -5 dB, -8 dB, and -11 dB. These S/N ratios were chosen based on extensive pre-testing that determined the level at which S/N participants scored 50% on the Modified Rhyme Test (i.e., -8 dB); thus, the other two S/N were chosen by adding (-5 dB) and subtracting (-11 dB) 3 dB from that level.

Dependent measures. The Modified Rhyme Test (MRT, discussed previously) was used to measure intelligibility in this experiment. The MRT consists of 300 monosyllabic English words grouped into 50 six-word sets. The sets of six words are arranged according to response ensembles, with each ensemble characterized by one vowel that is the nucleus of every word (ANSI, 1989). While most of the words are in the consonant-vowel-consonant (CVC) format, others are in the form of CV or VC.

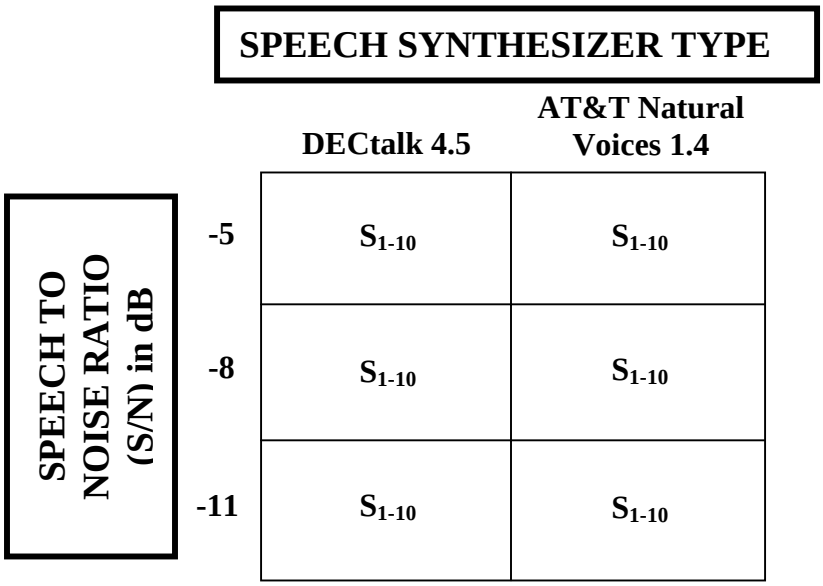


Figure 10. Experimental design for the speech intelligibility experiment.

A carrier sentence was used to present the participant with one word from each ensemble. The participant responded by circling one of the six words in each ensemble. It takes approximately two to three minutes to administer each 50-word set when carrier sentences are used. When carrier sentences are not used, the test takes approximately 75 seconds (ANSI, 1989) and, since the test has a closed-response set, it is easy to administer and score.

Apparatus

The simulator utilized in the study to produce aircraft engine noise, the i-GATE PC-ATD, is FAA-certified for pilot training (see Figure 1). It is a technically advanced digital training system designed around the ELITE training software and the 'smart' panel by ModWorks, Inc. and features the latest digital technology in support of creating a realistic environment. The simulator was used only to produce and present the aircraft engine noise of a Cessna 172R at 85 dB(A), which was measured as the cockpit noise level of the that aircraft's engine at cruise in a previous study (Lancaster et al., 2003).

The simulator was housed in the Human Factors Engineering and Ergonomics Center (HFEEC), in room 567 Whittemore Hall, on the Virginia Tech campus. The simulator is equipped with an 'experimenter's station' in the form of a personal computer located outside the testing room which is connected to and controls aspects of the simulation. Additionally, live video of the simulator room was captured using a Sony DXC-327 camera and was presented on a Sony Trinitron PVM-1341 monitor.

The simulator room has been lined with Sonex SCOC2 acoustical foam on all surfaces except the floor (carpeted) and ceiling (acoustical tile) in an effort to reduce the reverberation time to approximate the acoustics of a Cessna 172 cockpit (see Figure 11).

After application of the acoustical foam, measurement of reverberation resulted in the following times (*frequency, reverberation time in seconds):* 40 Hz, 0.4; 50 Hz, 0.45; 63 Hz, 0.4; 80 Hz, 0.28; 100 Hz, 0.11; 160 Hz, 0.39; 125 Hz, 0.4; 160, 0.39; 200 Hz, 0.19; 250 Hz, 0.18; 315 Hz, 0.11; 400 Hz, 0.13; 500 Hz, 0.18; 630 Hz, 0.16; 800 Hz, 0.15; > 1 KHz, all < 0.10. Realistic aircraft sounds produced by the simulator were channeled through a Parasound P/LD-100 line drive preamplifier and an OCM 200 Series amplifier and were presented through two (2) Infinity SM-155 loudspeakers at a sound pressure level (as mentioned) of 85 dB(A). Sound levels were verified before each session using a Larson-Davis 3200 spectrum analyzer. The spectrum analyzer was calibrated prior to each use using a Quest QC-20 calibrator to produce a 1000 Hz tone at 94 dB.

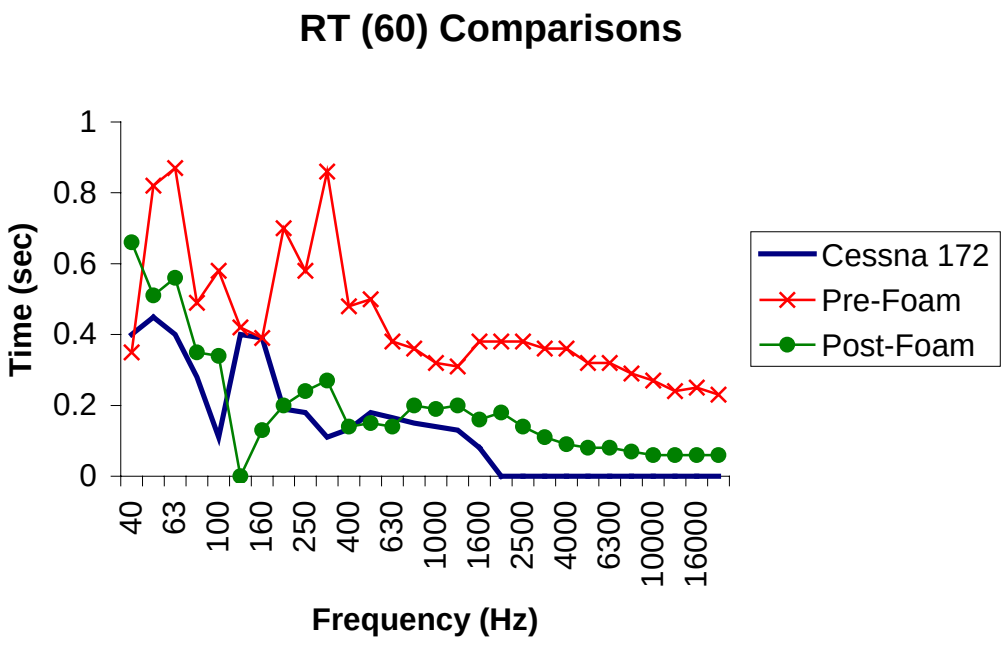


Figure 11. RT(60) results before and after foam application to the experimental chamber.

Procedure

Participants reported to the simulator laboratory. Before conducting any portion of the experiment, including pre-testing, all participants were required to read and sign an informed consent form (see Appendix B).

All MRT words (see Appendix C) were produced by each speech synthesizer within the carrier sentence "Mark the word _____ now." The speed of utterances was set at 60% for both TTS engines. This speed was determined through pre-testing to produce a sentence flow similar to that of recorded ATC directives. All manipulations of the synthesized words and sentences were conducted utilizing Cool Edit Pro v1.2a, and were verified using a Larson-Davis 3200 spectrum analyzer. The equalization process produced 600 sentences (300 from each synthesizer); these sentences were then randomly assigned to one of six sentence lists for each synthesizer. After all 300 sentences were recorded, the peak A-weighted level (using a "slow" [i.e., one second] meter time constant) for each word in each sentence was set as equivalent. To do so, the audio output of the computer was connected to the analog input of the spectrum analyzer and the amplification of each word in each sentence was digitally adjusted so that the levels of all utterances were equal. A two-minute segment of speech-weighted noise, generated by a Beltone 2000 audiometer, was also digitally recorded and its unweighted L_{eq} equated to the same level used to equalize the individual MRT words. [The peak A-weighted level of individual words measured using a slow response is approximately equal to the long-term unweighted L_{eq} of continuous speech (Kryter, 1985).] This relationship and the two-minute segment of speech-weighted noise allowed the speech levels to be quantified. The sentences were then saved as 44.1 KHz, 16-bit wav files. Each of the six 50-

sentence sets was assembled by concatenating individual sentences together in the desired order, with five seconds of silence between sentences. Sentence sets were then burned to separate compact discs. Speech stimuli were presented to the subjects through a Bose Active Noise-reduction (ANR) aviation headset (Model AHX-02) using a Sony CDP-XE400 compact disc player and a Sony STR-DE135 stereo receiver. As the current research seeks to investigate advanced technologies and their effect on pilot performance, it was thought prudent to include the state-of-the-art Bose headset instead of a traditional passive device (e.g., David Clark aviation headsets). The headset was set to 'mono' mode and the headset volume was adjusted and locked to 'maximum.' Speech levels were set by placing the headset on an acoustical test fixture containing a one-inch precision measurement microphone (Larson-Davis model 2575, serial # 1280) and adjusting the speech level using the receiver's volume control (see Figure 12).

Figure 12. Active noise-reduction aviation headset positioned on an acoustical test fixture.

Participant familiarization. Participants first practiced the MRT using real speech stimuli recorded by Auditory Systems Laboratory personnel. Participants were presented with six possible word choices on a response sheet while the target word, contained within the carrier sentence "Mark the word _____ now," was presented through the Bose ANR headset. The participant was instructed to select the word heard by marking his/her choice on the response sheet. Participants completed the MRT procedure twice within each of three S/N levels: -3 dB, -6 dB, and –9 dB. Background aircraft engine noise was produced by a Cessna 172R flight simulator (FlyELITE 'i-GATE') and was presented at 85 dB(A). This process was repeated until all 50 words were presented. The participant performed at least two practice trials of the MRT. Although the purpose of the practice trial was to familiarize the participant with the task, the participant's answers were examined to ensure that the participant was paying attention.

Data collection. After familiarization with the MRT procedures, each participant completed the MRT at the three S/N (-5 dB, -8 dB, and -11 dB) with each synthesizer tested on different days. Background aircraft engine noise was produced by a Cessna 172R flight simulator (FlyELITE 'i-GATE') and was presented at 85 dB(A). Participants were not present when the experimenter set, adjusted, and verified the speech level under the headset to obtain the S/N ratio under test. As in the familiarization session, participants were instructed to circle or otherwise mark each stimulus word heard over the headset. Each MRT was completed twice, using different sentence sets, to produce an average MRT score, per ANSI S3.2-1989. The raw scores were then adjusted for chance or guessing using the algorithm outlined within the MRT standard (discussed

below). The resulting intelligibility scores were expressed as percentages, ranging between 0 and 100%. After each set of trials with a particular speech synthesizer, participants were given a questionnaire to elicit their subjective impressions of each TTS engine. See the functional block diagram (Figure 13) for a schematic of the experimental setup.

Data analysis. The raw scores for the MRT were recorded into a separate text file and were copied into a Microsoft Excel spreadsheet. The raw scores were then entered into the following formula, from ANSI S3.2-1989, that adjusted the raw score for chance or guessing:

$$R_a = R - [W/(n-1)]$$

R_a is the number of items correct adjusted for chance/guessing, R is the number of items correct, W is the number of items incorrect, and n is the number of alternative choices per item.

The scores for the MRT were expressed in percentages and could range between 0 and 100%. However, a score of "0" would indicate that the participant was not paying attention; those participants were eliminated from the experiment. Once the R_A value was calculated, it was divided by 50 (the total number of words presented to the participant) and then multiplied by 100 to obtain the percentage of correct words. Appropriate statistical tests, including analysis of variance (ANOVA), were then conducted on the resulting *percentages correct* to determine whether a significant difference existed between the two speech synthesizers.

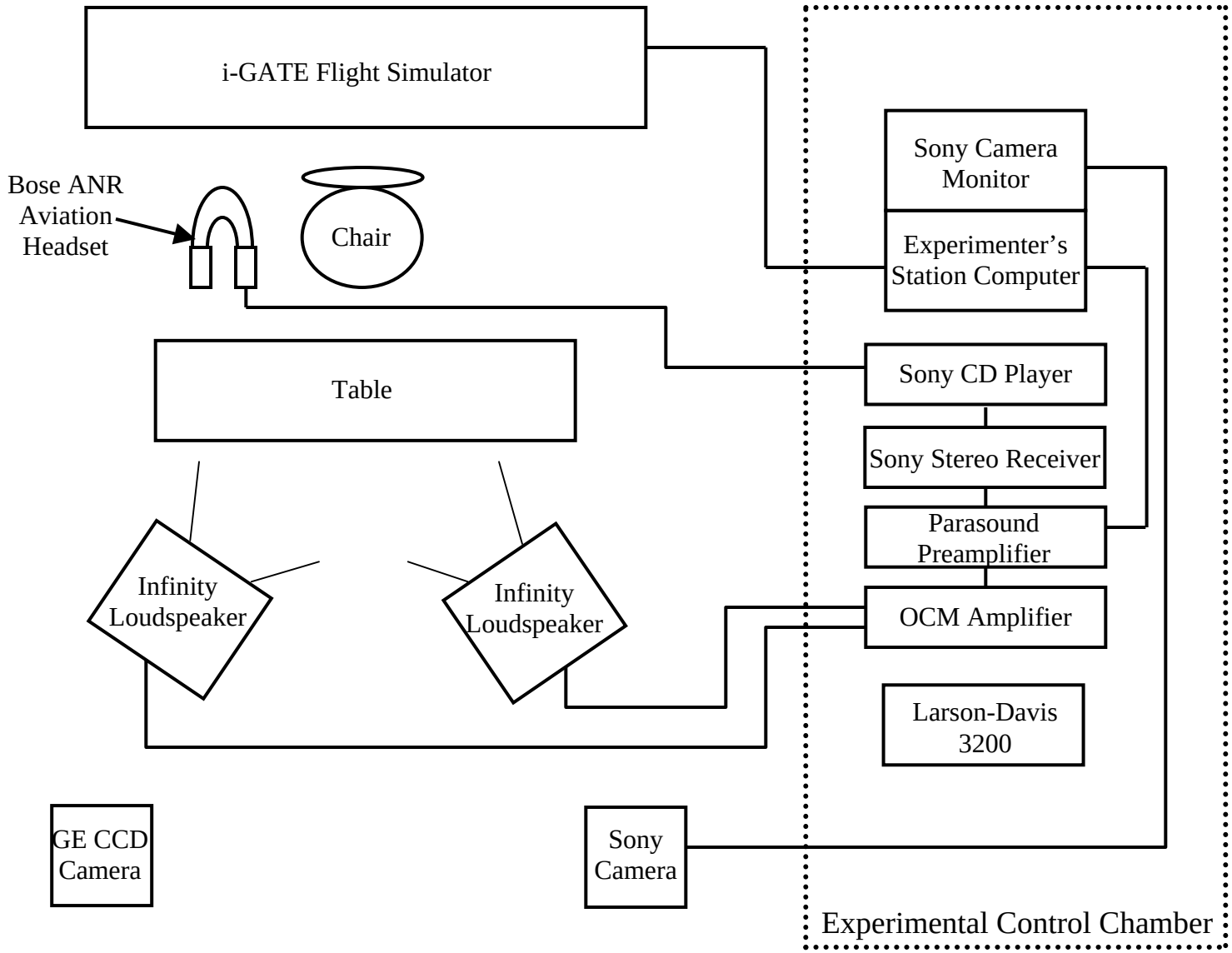


Figure 13. Functional block diagram of the experimental setup for the speech intelligibility experiment (not to scale).

RESULTS AND DISCUSSION: SPEECH INTELLIGIBILITY (Experiment I)

Main effects of TTS engine and S/N ratio. Analysis of variance (ANOVA) revealed a significant main effect of TTS engine on intelligibility collapsed across all speech-to-noise (S/N) ratios, $F(1,45) = 46.03$, $p < 0.0001$, with the AT&T product performing superior to the DECtalk product (see Table 13). Results are displayed in Figure 14. The mean intelligibility across all S/N ratios was 78.5% for the AT&T product and 59.6% for the DECtalk product. ANOVA analysis also revealed a significant main effect of S/N ratio on intelligibility $F(2,45) = 46.10$, $p < 0.0001$. Mean intelligibility for both TTS engines at –5 dB S/N, –8 dB S/N, and -11 dB S/N was 81.9%, 74.7%, and 50.6%, respectively.

TABLE 13

Analysis of variance for the speech intelligibility experiment

Source of Variance	df	MS	*F*	*p*
Text-to-speech Engine (TTS)	1	5339.26	46.03	<0.0001*
Speech-to-Noise Ratio (S/N)	2	5346.71	46.10	<0.0001*
S/N X TTS	2	91.31	0.79	0.4613
S/N X TTS X Subject	45	115.98		

*indicates significant result ($p \leq 0.05$)

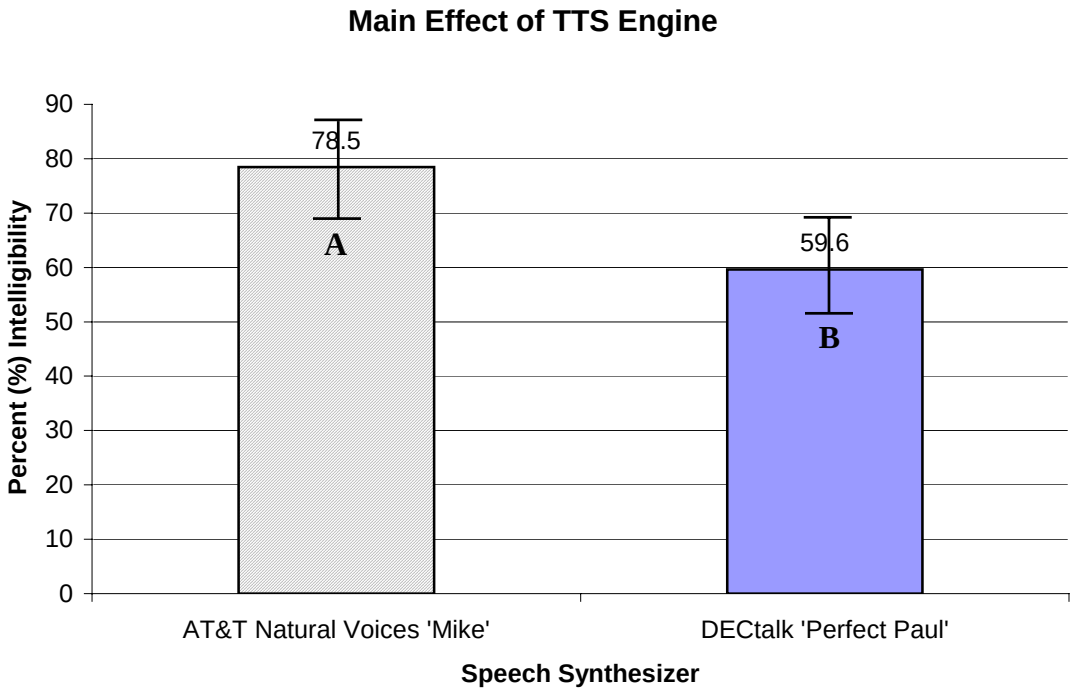


Figure 14. Main effect of TTS engine on intelligibility. Means with different letters are significantly different ($p < 0.05$).

Post-hoc analysis of percent intelligibility at the three speech-to-noise ratios using the Tukey's test (chosen because it is conservative with respect to alpha error) indicated significant differences in intelligibility between each S/N ($p < 0.05$); see Figure 15. Interaction analysis of S/N ratio by TTS engine was not significant. Linear contrast analysis revealed a significant linear trend, $F(1) = 84.46$, $p < 0.0001$. Mean intelligibility for the AT&T product was 90.2%, 86.4%, and 58.8% for -5 dB, -8 dB, and -11 dB S/N, respectively. Mean intelligibility for the DECtalk product was 73.7%, 62.6%, and 42.5% for -5 dB, -8 dB, and -11 dB S/N, respectively.

A significant linear trend indicates that the differences between the two TTS engines depend on the S/N ratio (see Figure 16).

Subjective impressions of the two speech synthesizers indicated a preference for the AT&T product's 'more natural-sounding voice.' Participants related that there 'was no comparison' between the two engines and that the DECtalk product 'sounded too mechanical.'

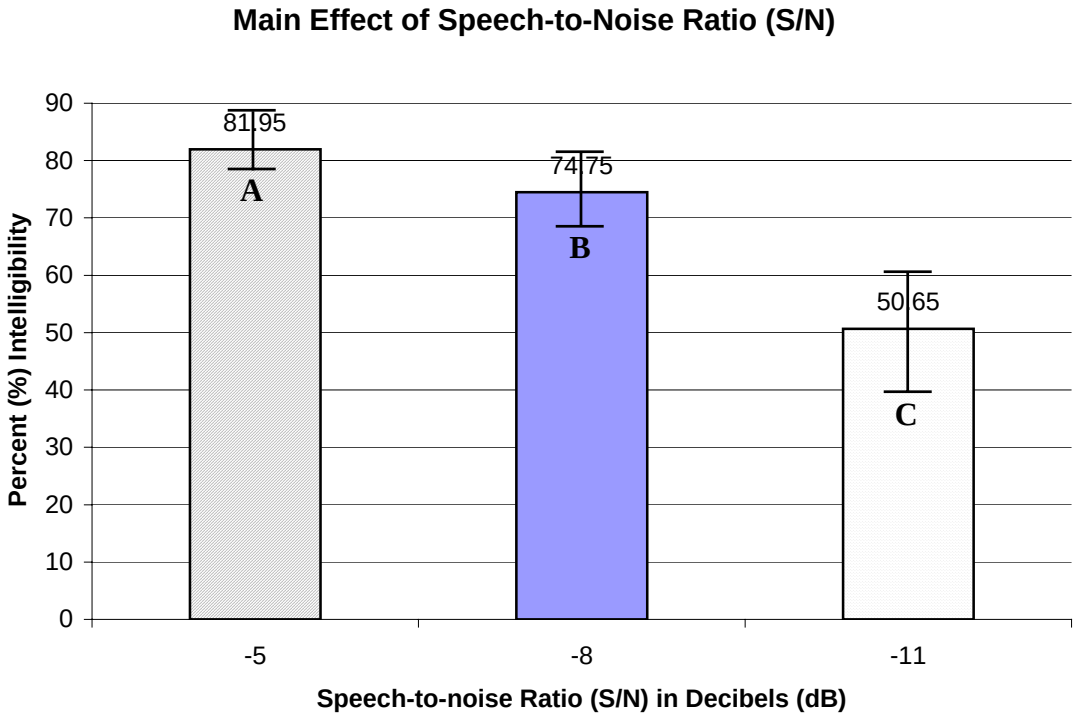


Figure 15. Main effect of speech-to-noise ratio (S/N) on intelligibility. Means with different letters are significantly different ($p < 0.05$) according to Tukey's test.

These results are not too surprising when one considers the aforementioned current prevalence of high-powered computing systems that foster the creation and iteration of realistic-sounding systems. This, of course, provides an advantage for the newer AT&T product. The DECtalk product, while having shown superior intelligibility

in the past, has not been updated recently and, in fact, the technology has since been sold by Digital Electronic Corporation as newer systems have been designed and tested. The AT&T product, which is a concatenative system (and thus uses recorded natural speech), sounds superior in its prosody. Still, utterances from the AT&T product can sometimes be 'choppy',' as was also related in subjective comments. Hopefully, iterations of the AT&T TTS engine will reduce this deficit.

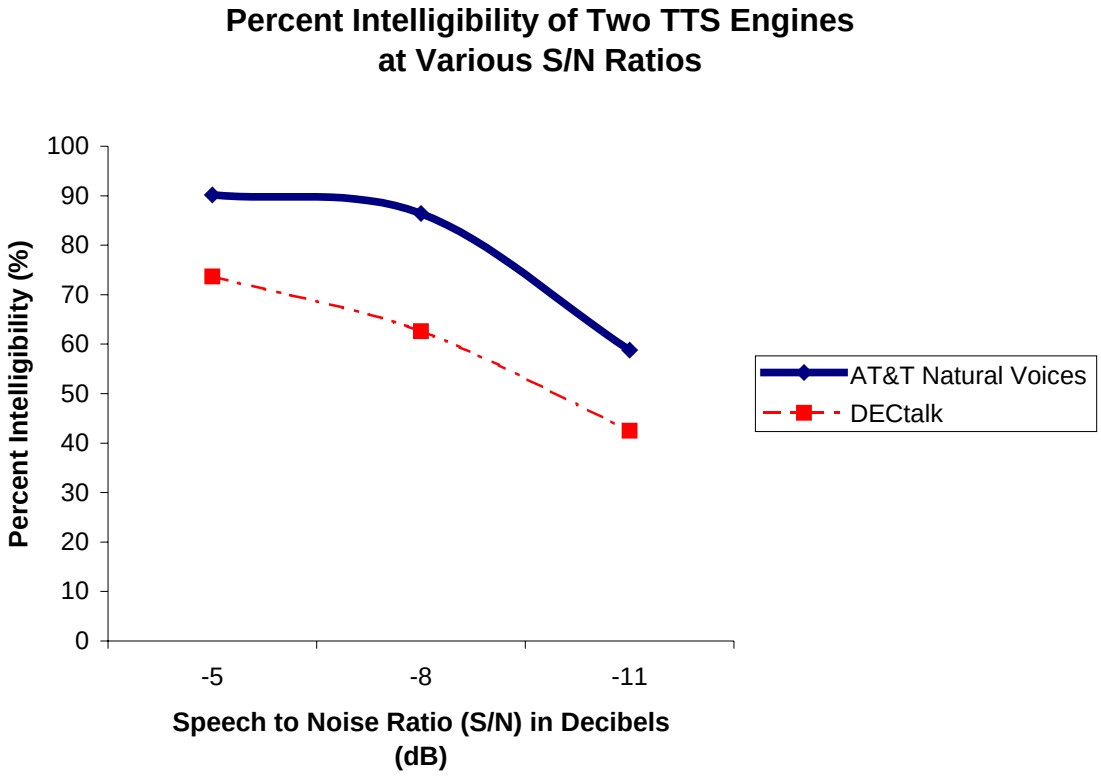


Figure 16. Linear contrast analysis indicates a significant linear trend ($p < 0.05$).

Conclusion. If synthesized speech is to be an option in future aviation systems that incorporate automation in the form of auditory displays, it is of utmost importance

that users can hear and understand the utterances produced by them in the prevailing noise environment. Indeed, in the proposed National Airspace System, in which automated systems will become increasingly prevalent in piloting operations (e.g., controller-to-pilot data link systems), verbal auditory displays must compete with many other existing (and future) auditory cues for the attention of the pilot (Rehmann 1996, 1997). Misunderstanding of clearance messages, vectoring commands, or traffic advisories can have severe implications within the aviation arena due to a host of factors, not the least of which is the speed of aircraft and the congestion in which they may operate. These concerns call for auditory displays that not only present at intensities that can be readily heard, but also, if a TTS engine is to be used, one that is as close to natural human speech as is possible should be chosen to foster intelligibility. As such, potential applications of this research include guidance for the integration of automated voice technologies in the cockpit and in similar systems that present elevated levels of background noise during normal communications and auditory display operations.

The results demonstrate that, even at S/N ratios which would not traditionally be recommended by human factors design principles, which advocate +10-15 dB S/N, the AT&T system still provided reasonable intelligibility (i.e., 82% at –5 dB S/N, 75% at –8 dB S/N). One must be careful in the application of heuristics such as these by taking into account the technologies that comprise the system. In this instance, the use of ANR headsets provides an appealing (from a subjective standpoint) and effective (from an objective standpoint) alternative to the traditionally passive devices worn by most pilots. As the prevailing noise presented in the current study was dominated by the largely low frequency spectrum produced from a piston-powered GA engine, the use of an ANR

headset is quite attractive given that its most effective attenuating capabilities reside within the low frequency range. Application of the aforementioned design principle of +10-15 dB S/N may actually result in a hearing hazard when using ANR, as signals presented over headsets that incorporate this technology likely do not require such levels.

METHODOLOGY: DATA LINK PERFORMANCE (Experiment II)

Experimental Design – Data Link Performance

There were two independent variables for this experiment, resulting in a 4 X 2 mixed design (see Figure 17). There were seven dependent measures. Each of these is described below. Before conducting any portion of the experiment, including pre-testing, all participants were required to read and sign an informed consent (see Appendix D).

Participants

Sixteen (16) current VFR-rated pilots were recruited to participate in the experiment. Every effort was made to recruit female pilots, but this effort was unsuccessful. As such, all participants were male. This sample represented the qualifications of pilots who may operate aircraft within envisioned future aviation systems. Pilots were recruited (through postings, email, listserv, and word-of-mouth) from the local Blacksburg community, Roanoke, and further outlying areas until the requisite number of participants was achieved. Pilot ages ranged from 21 to 70 with a mean of 38.5 years. Pilot flight time ranged from 97 to 5,500 flight hours, with a mean of 835.4 flight hours and a median of 282 flight hours (see Appendix E). Two of the 16 pilots related experience using textual data link in the 1970s.

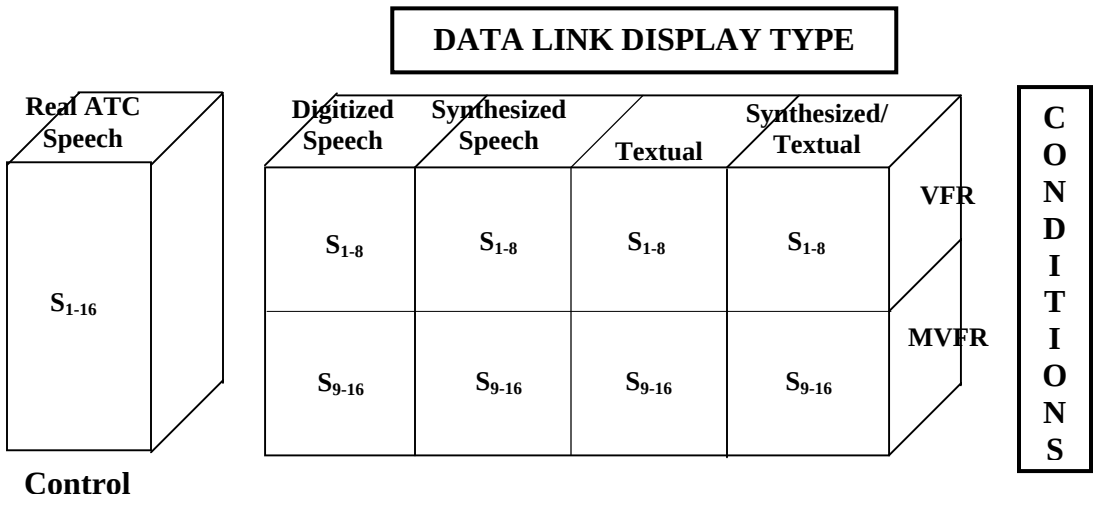


Figure 17. Experimental design for the mixed-modality simulated data link experiment.

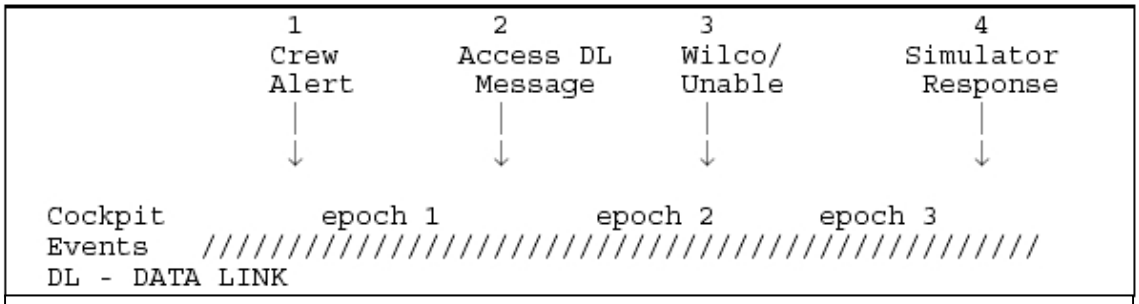
Independent measures. The first independent variable was representative of the presentation type of ATC information (a within-subjects variable with four levels). They are: textual, digitized speech, synthesized speech, and a synthesized/textual combination. The ATC commands presented within each were representative of typical control directives (e.g., "Cessna bravo two-seven, turn to heading 010, climb and maintain 6200 feet"), and were gleaned from actual recordings of radio traffic that were made in a previous experiment.

The second independent variable was representative of the flight condition (a between-subjects variable). The first condition was full visual flight rules (VFR), i.e., unlimited ceiling and visibility; the second ceiling level was in a marginal visual flight rules (MVFR) condition, i.e., clouds at 2800 feet above ground level (AGL) with three miles visibility. A control included real voice ATC presentation (i.e., not digitized but a live recording) within full VFR conditions. In order to reduce the effects of practice on

the experimental outcome, the treatment conditions were assigned by using a balanced Latin square.

Dependent measures. The seven dependent measures utilized in this experiment are described below.

The first three dependent measures specifically relate to temporal elements and are collectively called 'response time epochs.' These intervals correspond to those utilized in previous investigations of data link systems, and are diagrammed in Figure 18 (Rehmann, 1993, 1994, 1995). The epochs (1, 2, and 3) were logged based on key events in the data link transaction, were gleaned from videotape analysis of each experimental session, and were further supported through the simulator's timestamp feature.



Event 1: Incidence of crew alert announcing message pending
Event 2: Pilot presses message key to call message to viewing surface and/or voice presentation
Event 3: Pilot presses WILCO/UNABLE key
Event 4: Pilot makes control input and the simulator responds to uplink message

The time measurement epochs associated with each interval are:
Epoch 1: Begins with incidence of pilot data link auditory alert. Ends with pilot selecting the message for display.
Epoch 2: Begins with pilot selection of the message for display and ends with the selection of a WILCO or UNABLE response.
Epoch 3: Begins with a pilot WILCO or UNABLE selection and ends with a control input as indicated by navigation data transmitted by the simulator.

Figure 18. Key response time events. Adapted from Rehmann 1993, p. 11.

The fourth and fifth dependent measures related to workload. The fourth dependent measure was *head-up/head-down time*. Accessing, reading, and responding to data link messages may increase head-down time and, by extension, increase workload. In order to assess this, a video camera was positioned and aimed to show the back of the pilot's head as well as the instrument panel and a simulated control display unit (CDU) (a desktop computer, which served as the simulated 'data link' display); see Appendix E. During each flight, a video record was captured of pilot head movements. As each data link message was received (as marked by the activation of an auditory alert signal, discussed shortly), the time each pilot spent looking at the simulated data link CDU was measured using a digital stopwatch. It is understood that this procedure did not provide data on head-down time for reasons *other* than data link; however, it did allow an assessment of the amount of visual attention the data link system commands. Also as discussed, it is imperative that future aviation systems that incorporate increasing levels of automation support the need to maintain situation awareness as well as decreased workload. As hypothesized, the use of the textual display conditions should result in decreased head-up time related to the need to refer to the visual messages from ATC, thereby producing a negative effect on situation awareness, and will increase workload compared to the auditory-only conditions. The fifth dependent measure was therefore a subjective measure of workload determined using the MCH workload rating scale, which was administered after each experimental trial. As discussed previously, this particular scale has been validated for use in systems that incorporate automation and was thus the appropriate measure; it has also been utilized in several data link investigations in which simulators were used (Rehmann 1993, 1994; Rehmann, Reynolds, and Naumeier, 1993).

As discussed, there is concern within the aviation community that pilot SA will be degraded by the removal of the radio 'party line' as ATC communication shifts to data link. In order to evaluate this concern empirically, pilot SA was measured during each experimental run using two techniques: SAGAT and SART. Therefore, the sixth dependent measure was an objective measure of situation awareness using the *SAGAT*. As mentioned within the SA section specific to SAGAT, the probe queries utilized included those resulting from a GA SA requirements analysis that was conducted utilizing SMEs (see Figure 3) and from those identified by Rehmann (1993); see Table 7. Also as discussed, Endsley (2000) relates that SA probes should occur every two minutes, and even more often in an effort to improve sensitivity. This corresponded to twelve (12) probes (three sets of four queries) during each experimental trial (see the procedure section for the rationale behind this number). Questions were randomly chosen for probing based on the aforementioned SA GA requirements analyses but, with respect to probes inquiring into ATC transmissions, did not include those specific to transmissions which ATC has not yet made. The seventh dependent measure was a subjective measure of situation awareness determined using the *SART*, which was administered to pilots after each experimental run.

Apparatus

Please see the *apparatus* section of experiment 1 (speech intelligibility) for information regarding the flight simulator, sound equipment, and ancillary devices.

For the evaluation of data link, an auditory alerting signal indicated an incoming ATC message and thus began the timing interval for epoch 1. Rehmann (1993) used a SELCAL (selective callout) sound that is used in Boeing 727 aircraft, which is a two-

tone, mechanical doorbell chime. Even though post-hoc subjective evaluations in the Rehmann study resulted in a primary suggestion for changing the nondistinctive SELCAL sound to a sound that is more distinguishable from other cockpit aural indicators, it is argued here that its inclusion for this particular experiment was sufficient, as there were no other aural alerts within the i-GATE system that are similar to the SELCAL sound (Rehmann, 1993). Therefore, the SELCAL sound used for the current experiment was a digital wav file from a doorbell (Network Music, Inc.; Vol. 25, track #70 [first five seconds looped using Cool Edit Pro v1.2a]). This sound was similar to that used by Rehmann (1993), and was presented to the pilots through the Bose ANR aviation headset as were all verbal data link transmissions.

The Rehmann (1993) research also utilized a visual alerting signal – that of a 'blue aviation light', as an addition to the aural alerting sound. These two alerting mechanisms were then compared to see which, if any, were the most useful to the pilot teams. Given that GA typically involves a single pilot, and that future regimes such as SATS require a single pilot, it was thought sensible to include both alerting schemes (aural and visual) such that one provided redundancy to the other. No attempt was made to compare the two alerting schemes. As such, a 'dash light jumbo ¾ nut Blue' (Aircraft Spruce and Specialty Company part # 17-423) was installed on the i-GATE, just below the glare shield, to the right of the flat-panel display, and flush with the instrument panel (see Figure 1). See Figure 19 for a functional block diagram of the experimental setup.

Incoming data link ATC information was synthesized using the Fonix iSpeak system, which incorporated version 4.5 of the AT&T Natural Voices 'Mike' speech synthesis suite with the utterance speed set at 60%. The iSpeak system was installed on a

desktop computer with a touchscreen upon which all ATC commands were generated and through which all data link transactions were made (see Appendix E for touchscreen layout). Digitized ATC information was recorded using a Labtec Desk Mic 534 plugged into a Sound Blaster Live! audio card installed on a desktop PC. The synthesized and digitized sentences were then saved as 44.1 KHz, 16-bit wav files. The sounds were then uploaded to the simulator system and output (and/or digitized) was presented on the Bose ANR aviation headset. Implementation of SAGAT queries, which required the 'blacking out' of the flat panel display upon which all instruments are depicted (disallowing visual reference), was accomplished through an interruption of the video stream within the simulator itself using a Zonet keyboard-video-mouse (KVM) 3002 switch.

Procedure

Participants reported to the Auditory Systems Laboratory, 538 Whittemore Hall, on the Virginia Tech campus. Before conducting any portion of the experiment, including pre-testing, all participants were required to read and sign an informed consent document (see Appendix D).

Participant familiarization. Participants first completed an audiogram (see Appendix E) to determine their hearing level (HL). Participants were then introduced to and allowed to fly the i-GATE simulator.

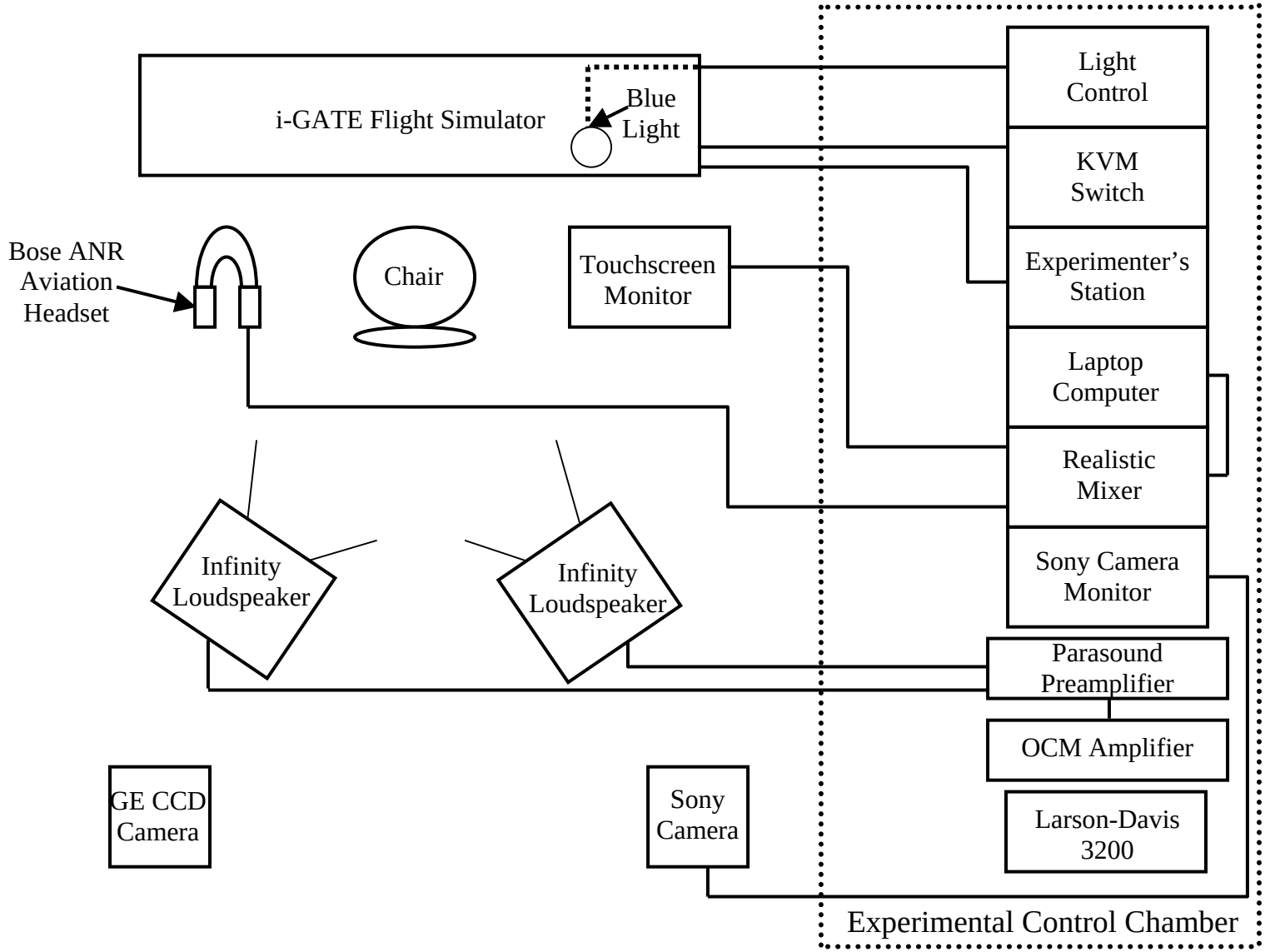


Figure 19. Functional block diagram of the experimental setup for the simulated data link experiment (not to scale).

The pilots flew a simulated daylight route in a fictitious airspace with unlimited ceiling and unlimited visibility. This procedure simulated a typical enroute flight scenario, with ATC relating vectors and other information to the pilot. A training criterion was considered reached after the pilot demonstrated command of the simulator as evidenced through successful responses to ATC directives (i.e., the pilot correctly responded to directives in a timely manner and could maintain control of the aircraft at all times). However, the pilot was allowed to continue the familiarization session for up to one hour if he/she desired more practice. If the pilot failed to meet the training criterion after one hour, he/she was dismissed from further participation.

The final task during familiarization was to have each pilot set his 'most comfortable hearing level' (MCL) of a synthesized speech passage under the ANR headset within aircraft engine noise at 85 dB (A). As a stimulus in this experiment included the superior-performing TTS engine from the speech intelligibility study (i.e., AT&T's Natural Voices 'Mike'), it was imperative that the hearing levels of the pilots 'matched' those of the participants in the previous study. To do so, consideration was given to Sanders and McCormick (1993) in their chapter on *Noise.* The following steps ensured equalization of hearing levels of the participants in both studies:

1. First, the audiogram threshold results obtained from the participants in the speech intelligibility study between 500 Hz and 2000 Hz were averaged, as were the minimum threshold results in that range. The audiogram threshold results of the pilots were treated similarly.

2. Next, the mean HLs of the pilots at 500 Hz to 2000 Hz were subtracted from the mean HLs of the participants in the speech intelligibility study. While listening to a passage of synthesized speech (i.e., AT&T's 'Mike' uttering a news story), the pilots were asked to set their MCL using the data link computer's volume control slider. The ANR headset was then placed on the acoustical test fixture (see Figure 12) and the MCL was measured and recorded using the Larson-Davis spectrum analyzer.
3. Recall that the SPL of the aircraft engine noise under the ANR headset was previously measured to be 64 dB(A). That number was added to the absolute value of the difference between the mean HL of the pilots subtracted from the mean HL of the speech intelligibility participants to produce an intensity that would result in a 0 dB SPL under the headset.
4. If the pilot's measured MCL was *less* than the intensity required to produce 0 dB under the headset, a 'correction factor' was applied to the intensity of the speech stimuli to produce 0 dB. The correction factor was the arithmetic difference in decibels between the MCL and 0 dB. Thus, the intensity of the synthesized speech stimuli in the data link study was standardized to match the intensity of the stimuli in the speech intelligibility study.

***Data collection*.** Pilots flew four routes (one for each data link condition, see Appendix E) within simulated class B airspace over central Florida. This region was chosen due to its lack of terrain (i.e., hills or mountains) that could have adversely affected pilot performance. The presentation order of the data link conditions was

counterbalanced, with two routes flown over two days. Each flight route included 14 different interactions (i.e., data link messages) with the simulated data link system. These messages included ATC directives (e.g., heading change), traffic advisories, ATIS messages, and expedited commands (i.e., emergency or time-critical maneuvers). Weather, malfunctions, and system failures were not enabled. The flight performance data from each participant's flight scenario (such as heading or airspeed for SA measures) were automatically collected by the simulation software and written to a local file on the experimenter's computer. Of these variables, the following indices were of interest: heading, altimeter, and indicated airspeed (IAS).

All interactions with the system (flight simulator and data link touch screen) were recorded on VHS videotape using the GE CCD video camera. The camera was positioned such that the pilot's head, flight simulator, and data link touch screen were clearly visible. Time spent within each epoch was determined through videotape analysis using a digital stopwatch.

Twelve SAGAT queries were introduced during each flight (one set of four queries approximately every five minutes). During the queries, the i-GATE monitor was 'blacked out' (using the KVM switch) such that no information could be gleaned from the pilot about the aircraft state. The three indexes of heading, altimeter, and IAS were used for part of the SAGAT determinations (i.e., was the pilot correct or incorrect in their response? [see the SAGAT section for examples of queries to be used]). These items were scored immediately upon exiting the experimental chamber and before the flight trial was resumed. Responses to other queries, such as the last ATC command heard, the location of last known traffic position, or frequency changes were recorded to determine

a correct or incorrect response. The GE CCD video camera recorded all experimental runs. Time spent in a 'head-down' condition was determined through post-experiment videotape analyses using a digital stopwatch.

In addition to flight and data link performance variables, subjective data was collected: workload was assessed using the MCH scale and subjective situation awareness was assessed using the SART. Additionally, a questionnaire eliciting pilot impressions of the data link modalities was administered (see Appendix F). These measures were obtained at the completion of each experimental trial.

***Data analysis*.** The i-GATE data collection module collected all data pertaining to flight performance measures, and the videotape analysis produced the epoch data. Appropriate statistical measures including ANOVA were computed on some of the resulting sets of dependent measures using the Statistical Analysis Software (SAS) program. An assumption was made with respect to the *response time epochs*—that they exist as separate, discrete temporal elements. It is for this reason that a multivariate analysis (i.e., MANOVA) was not indicated and separate, univariate procedures were conducted. Other tests were substituted as appropriate (e.g., Fishers' Exact Test). Fisher's tests were conducted on the MCH and SAGAT because the data represent nonparametric indexes (categorical responses and binomial responses, respectively). A probability value (p) of 0.05 was chosen as the cutoff level for statistical significance. These measures provided data on the amount of variation within each subject for display type, flight condition, and order of flight condition, and whether any observed differences between means may be due to chance or to systematic differences among the population

means. The analyses for each condition tested for significant differences in the dependent measures and subjects for both within and between elements. Post-hoc, unplanned comparisons of any significant *F* ratios revealed by the ANOVA analyses were conducted to isolate significant main effects and interactions of the overall ANOVA using Tukey's test. Tukey's was chosen because it is conservative with respect to alpha error, it conducts *all* pair-wise comparisons, and is a widely accepted post hoc measure.

RESULTS AND DISCUSSION: DATA LINK PERFORMANCE (Experiment II)

Workload

Workload ratings (Modified Cooper-Harper scale). Workload ratings were classified into three categories consistent with the Modified Cooper-Harper (MCH) Scale:

- 1-3 = acceptable level of workload
- 4-6 = high workload, should be reduced
- 7 and higher = major deficiencies, system design strongly recommended

Fisher's Exact test was appropriate for this data set due to the MCH's nonparametric, categorical ratings. Fisher's Exact test indicated that the four experimental data link conditions did not have a uniform workload perception ($p < 0.0001$). Results are displayed in Table 14. Figure 20 displays this non-uniformity of workload perception. The MCH scale groups workload rankings into discrete categories. These categories are 'acceptable level of workload' (1-3), 'high workload' (4-6), 'major design deficiencies

with redesign recommended' (7-9), and 'major design deficiencies with mandatory redesign' (10). Since none of the 16 pilots gave a rating of 10 for any of the conditions, this category is not displayed in Figure 20. The non-uniformity of workload perception is most notable within the textual data link condition. For the other three data link conditions, which represent verbal data link presentations or a combination of verbal/text, none of the 16 pilots reported a workload rating of (7+).

TABLE 14

Contingency table, conditions vs. categories of workload perception as measured using the MCH workload rating scale

Frequency *Percent* *Row Percent* *Col Percent*	Acceptable Level of Workload (1-3)	High Workload (4-6)	Major Deficiencies (7+)	Total
Textual	3 4.69% 18.75% 6.98%	10 15.63% 62.50% 55.56%	3 4.69% 18.75% 100.00%	16 25.00
Synthesized & Textual	14 21.88% 87.50% 32.56%	2 3.13% 12.50% 11.11%	0 0.00% 0.00% 0.00%	16 25.00
Digitized	12 18.75% 75.00% 27.91%	4 6.25% 25.00% 22.22%	0 0.00% 0.00% 0.00%	16 25.00
Synthesized	14 21.88% 87.50% 32.56%	2 3.13% 12.50% 11.11%	0 0.00% 0.00% 0.00%	16 25.00
Total	43 67.19	18 28.13	3 4.69	64 100.00

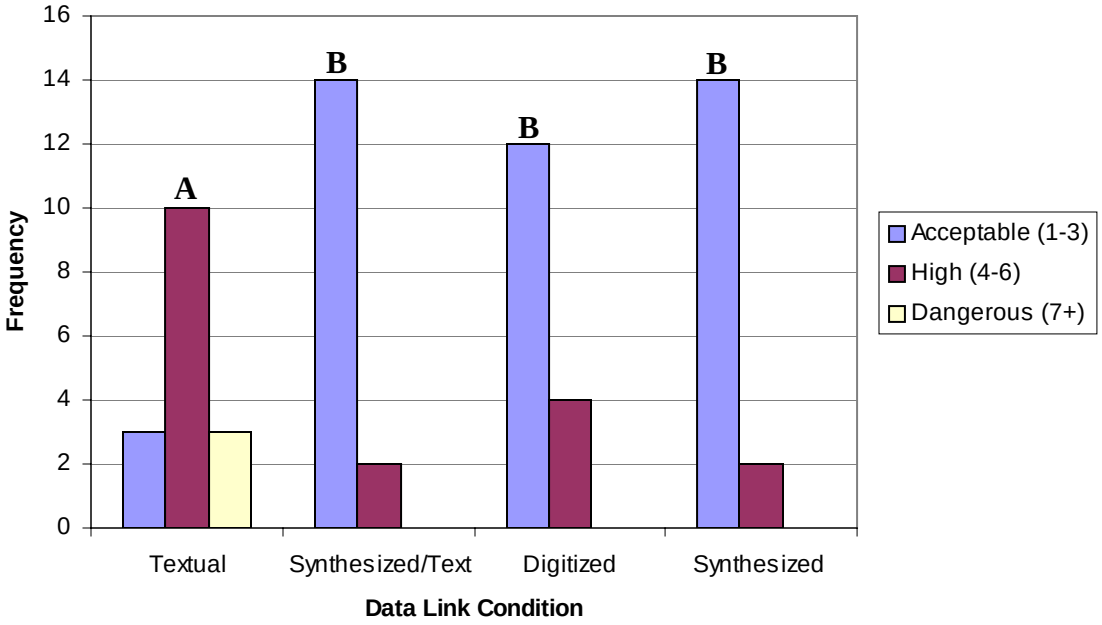


Figure 20. Workload perception across data link conditions as measured using the MCH workload rating scale. Distributions with different letters are significantly different ($p < 0.05$) according to Fisher's Exact test.

However, for the textual data link condition, three pilots gave a workload rating of (7+). Additionally, only three pilots reported workload as 'acceptable' (1-3) within the textual data link condition. Both data link conditions that incorporated synthesized speech resulted in the same workload ratings, suggesting that workload within these two conditions is uniform. For the digitized data link condition, four pilots reported workload as high (4-6). Thus, the subjective workload results from this study suggest that the textual data link condition results in high workload, and may not be an appropriate solution with respect to data link implementation. When rerunning Fisher's test to compare two conditions at a time, the textual data link was significantly different from the other data link conditions ($p < 0.0001$). The other three data link conditions were not significantly different.

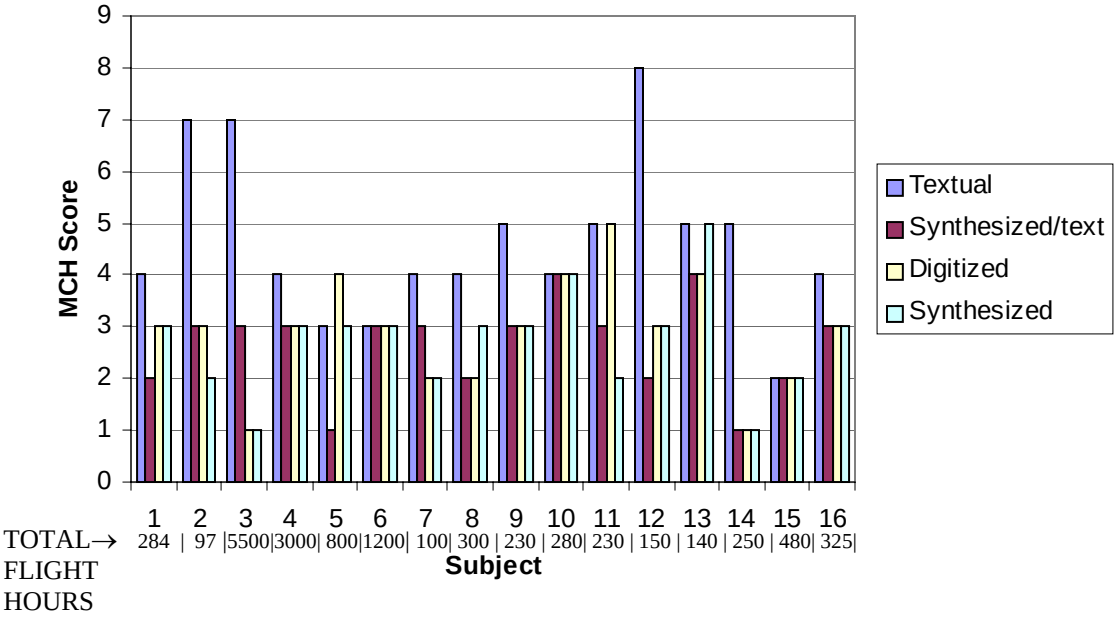


Figure 21. Workload ratings for each pilot as measured using the MCH workload rating scale.

Figure 21 shows the MCH raw scores for each of the sixteen pilots. This graph reveals that subjects 2, 3 and 12 perceived workload as unacceptably high (7+) for the textual data link condition whereas subjects 1, 4, 6, 10, 13, 15, and 16 perceived a uniform level of workload across conditions. Pilot flight experience did not seem to influence workload perception as subjects 3 and 4 were the most experienced pilots (average flight hours 4,250) but had differing workload perceptions, especially for the textual data link condition.

Workload perception appears to be acceptable for both synthesized and synthesized/textual data link conditions. Only two pilots rated the synthesized speech condition as 'high workload.' However, four pilots rated the digitized speech condition as 'high workload,' and ten of the pilots rated the textual conditions as 'high workload.' The textual data link was the only condition rated as 'unacceptably high' (three pilots).

It was interesting that the subjective workload ratings were not significant between flight conditions (i.e., VFR vs. MVFR). As MVFR rule maintenance requires constant monitoring and control actuation with respect to cloud separation distance, it was expected that the MVFR pilots would report increased workload when compared to the VFR pilots, but they did not. This result is likely due to the limitations of the simulator itself: views 'outside' were limited to a *very* small (1.5 inch) part of the display screen at the top. Perhaps there simply were not enough visual cues with respect to the outside environs with which to maintain separation and thus affect perceived workload.

***Workload: head down time*.** An objective measure of workload can be gleaned from observations of the time each pilot spent in a 'head down' condition; that is, not looking out of the simulated windscreen. Videotape analysis of all experimental conditions resulted in the total time (in seconds) spent 'head down.' ANOVA for these times revealed a significant effect of data link condition on head down time, $F(3,871) = 22.89$, $p < 0.0001$, with the means ranging from 9.08 seconds to 13.49 seconds (see Table 15). Results are displayed in Figure 22.

The mean amount of time spent 'head down' in the textual, synthesized, digitized, and synthesized/text data link conditions was 13.49 seconds, 9.13 seconds, 9.64 seconds, and 9.08 seconds, respectively. Post hoc analyses using Tukey's test revealed significant differences between the textual data link condition and the other three conditions (i.e., synthesized, digitized, synthesized/text). The other three conditions were not found to be significantly different. Head down time in the textual data link condition was significantly increased from head down time in the digitized condition, $t(871) = -6.15$, $p < 0.0001$.

TABLE 15

Analysis of variance for head down time

Source of Variance	df	MS	*F*	*p*
Between				
Flight Condition (FC)	1	4.31	0.60	0.4507
Subjects (S/FC)	14	7.19		
Within				
Data Link Condition (DL)	3	1004.66	22.89	< 0.0001*
FC X DL	3	27.65	0.63	0.5858
DL X FC X S/FC	871	43.89		

*indicates significant result ($p \leq 0.05$)

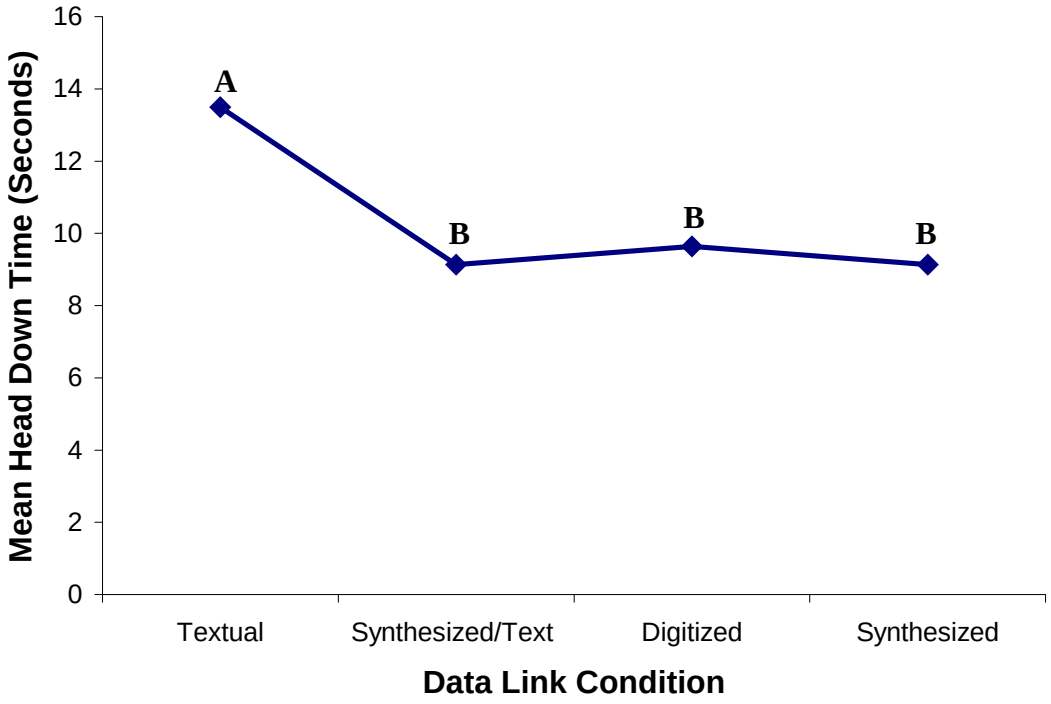


Figure 22. Mean head down time across data link conditions. Means with different letters are significantly different ($p < 0.05$) according to Tukey's test.

The textual condition also resulted in significantly increased head down time when compared to the synthesized condition, $t(871) = -6.96$, $p < 0.0001$. Finally, the textual

condition resulted in significantly increased head down time when compared to the synthesized/text condition, $t(871) = -7.04$, $p < 0.0001$.

Evaluation of the subjective workload ratings (i.e., MCH) and the objective workload evaluations (i.e., head down time) reveal many similarities. The most evident of these is in the textual data link condition. Recall that the MCH scores indicated that the workload perception of the textual condition was significantly higher than the other three conditions; similarly, the head down time analysis within the textual condition mirrored this effect (i.e., with longer observed times away from the out-the-windshield scene). Thus, the objective workload ratings support and reinforce the subjective workload ratings. Head down time is an important measure because any time spent 'head down' is time spent away from monitoring not only other cockpit systems (e.g., artificial horizon, airspeed indicator) but also the outside world (e.g., other aircraft, weather). A practical example of this can be seen in a simple equation for time, distance, and rate:

$$\boldsymbol{D = R \times T}$$

Consider the same aircraft used in this study, the Cessna 172R. Assuming an average speed of 110 knots, with no wind or other conditions that may affect speed, the distance traveled while head down in the textual condition (average 13.49 seconds) would be 0.41 nautical miles (nm), which equates to 0.47 miles, or about half a mile. Contrast this with the condition that resulted in the least average head down time (synthesized/text, 9.08 seconds), which results in 0.27 nm, or 0.31 miles traveled. The difference in distance may not seem like much, but for a fast-moving aircraft operating in a congested airspace, it could be the difference in successfully avoiding a potential conflict.

Data link performance

Epoch analysis. Each data link transaction was evaluated with respect to the time required for pilots to access (epoch 1), respond (epoch 2), and make a control input (epoch 3) with respect to the data link message. These evaluations were accomplished through videotape analysis using a digital stopwatch for each of the three epochs. During the videotape analysis, and with respect to epoch 3, it became apparent that pilots would routinely initiate a control input to the simulator before the completion of epoch 2, similar to the results obtained in the data link research of Rehmann (1996, 1997). Control input before epoch 2 completion was especially apparent for *expedited* data link messages, which will be discussed later. Such activity resulted in 'zero' times in epoch 3 for a vast majority of data link transactions, thereby disallowing a cogent or useful analysis of this time frame. Therefore, it was decided to remove epoch 3 data from the analysis.

Recall that epoch 1 represented the time required to access the data link message after the activation of the alerting stimuli (i.e., glare shield light and bell chime). ANOVA for the epoch 1 data revealed a significant interaction between flight condition and data link condition, F (3, 871) = 5.83, p = 0.0006, with the means ranging from 2.59 seconds to 3.69 seconds (see Table 16). Results from Tukey's test are displayed in Figure 23.

TABLE 16

Analysis of variance for epoch 1 time

Source of Variance	df	MS	*F*	*p*
Between				
Flight Condition (FC)	1	0.11	0.08	0.7876
Subjects (S/FC)	14	1.42		
Within				
Data Link Condition (DL)	3	36.17	14.40	< 0.0001*
DL X FC	3	14.64	5.83	0.0006*
DL X FC X S/FC	871	2.51		

*indicates significant result ($p \leq 0.05$)

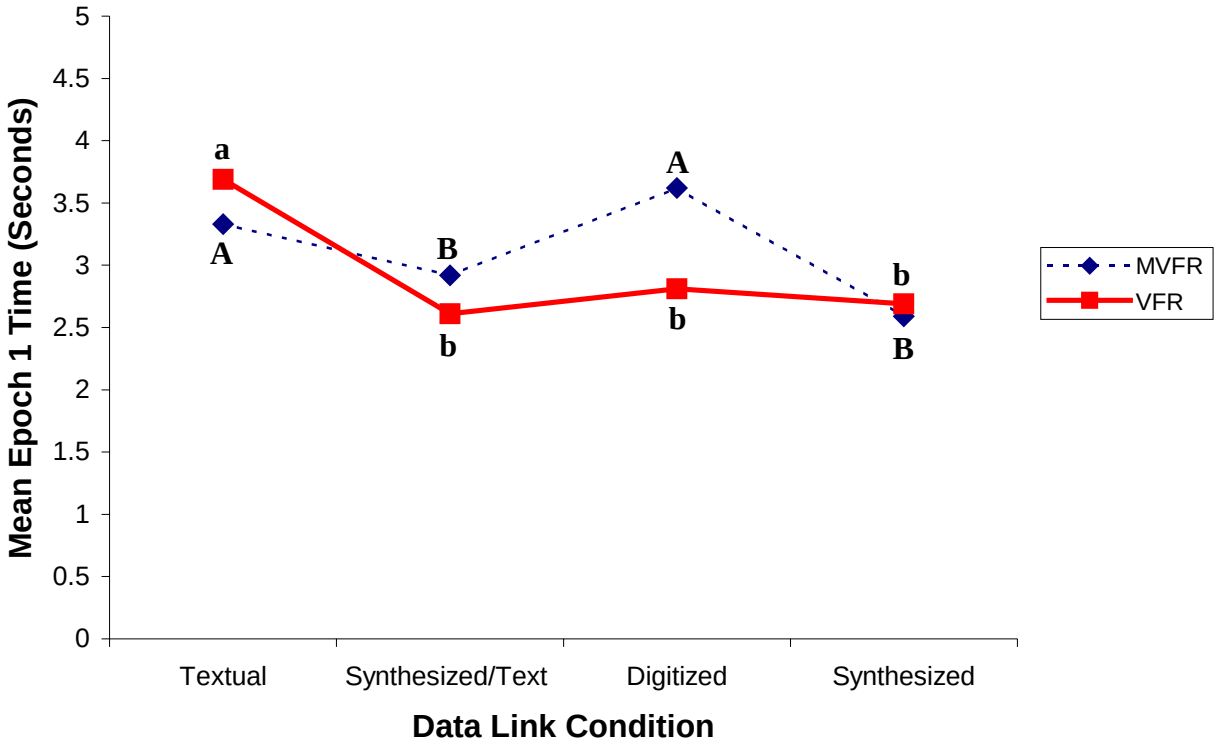


Figure 23. Interaction effect of epoch 1. Means with different letters are significantly different ($p < 0.05$) according to Tukey's test.

Having a significant interaction within this data set means that the effects of the four data link conditions with respect to epoch 1 varied between the VFR and MVFR flight conditions. Therefore, data link conditional effects are not independent of flight condition conditional effects. The most useful strategy with which to examine and interpret how the interaction means are affected is to look at a 'slice' of the flight conditions to see how the data link conditions varied between VFR and MVFR. Within the VFR flight condition, the textual data link condition resulted in significantly longer epoch 1 times than the other three data link conditions, which did not differ among themselves. Post hoc analysis using Tukey's test revealed that the textual data link condition resulted in increased epoch 1 time when compared to the digitized data link condition, $t(871) = -4.18$, $p = 0.0008$. The synthesized data link condition was found to result in decreased epoch 1 time when compared to the textual data link condition, $t(871) = -4.73$, $p < 0.0001$. Similarly, the synthesized/textual data link condition resulted in decreased epoch 1 time when compared to the textual data link condition, $t(871) = -5.09$, $p < 0.0001$. However, within the MVFR flight condition, the digitized data link condition ceases to differ from the textual data link condition. The synthesized data link condition took less epoch 1 time than did the textual data link condition, $t(871) = -3.48$, $p = 0.01$. The synthesized data link condition was found to result in less epoch 1 time than the digitized data link condition, $t(871) = 4.84$, $p < 0.0001$. Finally, the synthesized/textual data link was similarly found to differ from the digitized data link condition, $t(871) = 3.29$, $p = 0.02$.

The differences revealed within epoch 1 are somewhat surprising. As the alerting stimuli (i.e., bell chime and blinking light) did not differ across conditions, differences in

epoch 1 time would not be expected. But the fact that the data link conditions incorporating synthesized speech resulted in shorter epoch 1 times than both the digitized and textual data link conditions within the MVFR data set as well as from text in the VFR data set may suggest that some element of 'expectation' for synthesized speech exists. The pilots were made aware of the data link condition for each trial immediately before the trial began, so there could be no appreciable pre-trial bias with respect to data link presentation. However, the fact that the largest difference *between the means* with respect to epoch 1 time was less than one second overall (0.81 seconds) suggests that the observed differences between these means, across both data link conditions and between flight conditions, are not practically useful.

Analysis of epoch 2 data revealed a significant main effect of data link condition, $F(3,871) = 7.15$, $p < 0.0001$, with the means ranging from 14.58 seconds to 19.18 seconds (see Table 17). Results are displayed in Figure 24. The mean amount of time spent within epoch 2 for the digitized, synthesized, textual, and synthesized/textual data link conditions was 15.42 seconds, 15.28 seconds, 19.18 seconds, and 14.58 seconds, respectively. Post hoc analysis of epoch 2 data using Tukey's test revealed a significant increase in epoch 2 time between the textual data link condition and the other three data link conditions, which did not differ among themselves. The textual data link condition took longer than the digitized data link condition, $t(871) = -6.15$, $p < 0.0001$. The synthesized data link resulted in shorter epoch 2 time than did the textual data link condition, $t(871) = -6.96$, $p < 0.0001$. Finally, the textual data link resulted in longer times when compared to the synthesized/textual data link condition, $t(871) = -7.04$, $p < 0.0001$.

TABLE 17

Analysis of variance for epoch 2 time

Source of Variance	df	MS	*F*	*p*
Between				
Flight Condition (FC)	1	0.00	0.00	0.9529
Subjects (S/FC)	14	1.36		
Within				
Data Link Condition (DL)	3	944.58	7.15	< 0.0001*
DL X FC	3	33.02	0.25	0.8609
DL X FC X S/FC	871	132.11		

*indicates significant result ($p \leq 0.05$)

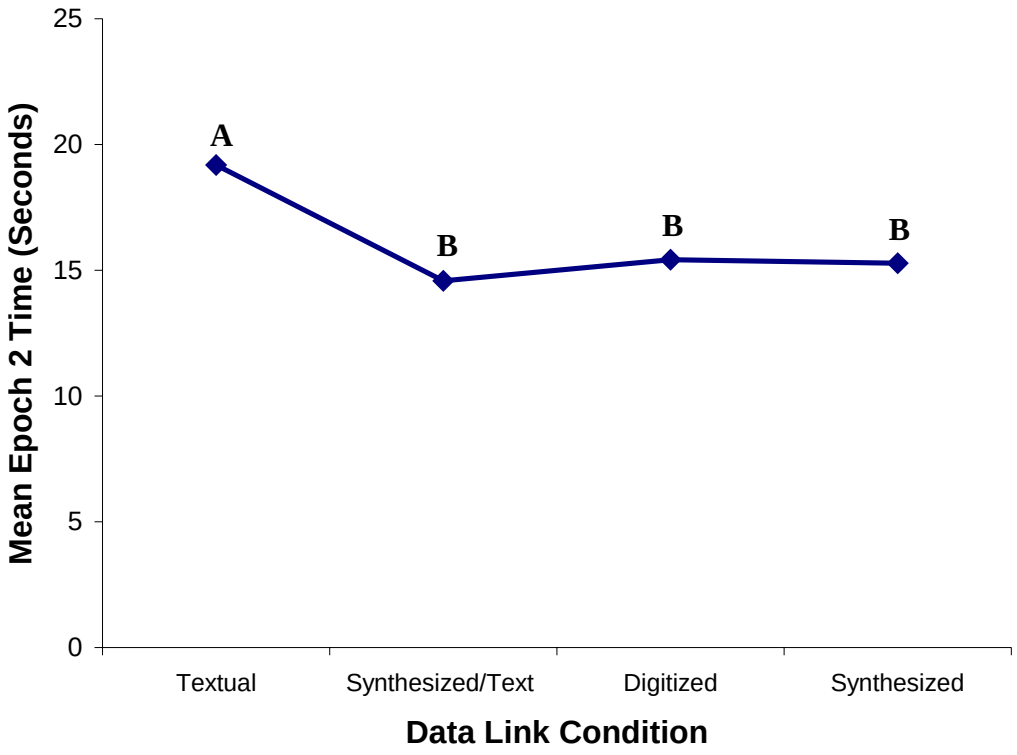


Figure 24. Mean epoch 2 time across data link conditions. Means with different letters are significantly different ($p < 0.05$) according to Tukey's test.

Epoch 2 represents the most interesting aspect of this study and the variation within it across data link conditions is quite compelling. As the modality of data link presentation was the main focus of the research, it was expected that the attention required within the textual data link condition would result in increased epoch 2 time. That is, the pilots were required to read the data link message, which existed on the data link touch screen, and the time required to complete that task would be increased over the time required to hear that same message aurally. The textual data link condition clearly took *longer* for pilots to complete (4.1 seconds) than any of the other three data link conditions, which varied less than a second between them. When considering the distance, rate and time formula discussed earlier, the increased time required to respond within the textual data link condition equates to 0.58 nm, or 0.68 miles traveled in that time. Contrast this with the data link condition that resulted in the shortest epoch 2 time (synthesized/text) which equates to 0.44 nm, or 0.51 miles traveled in that time. Of course, other concerns exist with respect to the effects of these data link modalities, such as the previously discussed affect on workload. The increased perceived workload, increased head down time (objective workload), and increased epoch 2 time all point to a serious concern with respect to a textual-only data link presentation.

Expedited commands. Within each flight trial there were two data link messages that corresponded to 'expedited commands'; that is, messages that required an immediate action or timely response from the pilot. Expedited command 1 was *"Cessna 519, expedite an immediate right turn to heading 'xxx', vector away from traffic."* Expedited command 2 was *"Cessna 519, immediate left turn heading 'xxx'."* With respect to the textual and synthesized/textual data link conditions, the textual elements of the expedited

messages were in **bold** and in ALL CAPS. While the visual presentation of these commands were not varied (e.g., backlit words vs. boldfaced words), it was thought prudent to try to convey the immediacy of the textual message through some format; therefore, bold and all caps was chosen. Due to the constraints of the TTS engine, there was no facility with which to stress the utterance of the expedited commands; therefore, to ensure equality across the aural data link conditions, the digitized expedited command utterances were not stressed either. Both expedited commands were analyzed for any effects between flight condition, across data link condition (i.e., within epochs 1 and 2), and in head down time.

Analysis of expedited command 1 data for epoch 2 revealed a significant main effect of data link condition, $F(3,42) = 3.03$, $p = 0.04$, with the means ranging from 10.65 seconds to 14.98 seconds (see Table 18). Results are displayed in Figure 25. Post hoc analysis using Tukey's test indicated a significant decrease in expedited command 1 performance between the synthesized data link and textual data link conditions, $t(42) = -2.82$, $p = 0.04$. The other three data link conditions did not differ among themselves with respect to expedited command 1 time.

The synthesized speech data link condition resulted in reduced expedited command 1 time within epoch 2. It is difficult to pinpoint the reasoning behind this result. Decreased expedited command 1 time might be expected to repeat within the synthesized/textual data link condition, as that condition maintains the same auditory component, but it did not.

TABLE 18

Analysis of variance for expedited command 1 time, epoch 2

Source of Variance	df	MS	*F*	*p*
Between				
Flight Condition (FC)	1	12.65	3.45	0.0843
Subjects (S/FC)	14	3.66		
Within				
Data Link Condition (DL)	3	57.14	3.03	< 0.0396*
DL X FC	3	8.10	0.43	0.7344
DL X FC X S/FC	42	18.85		

*indicates significant result ($p \leq 0.05$)

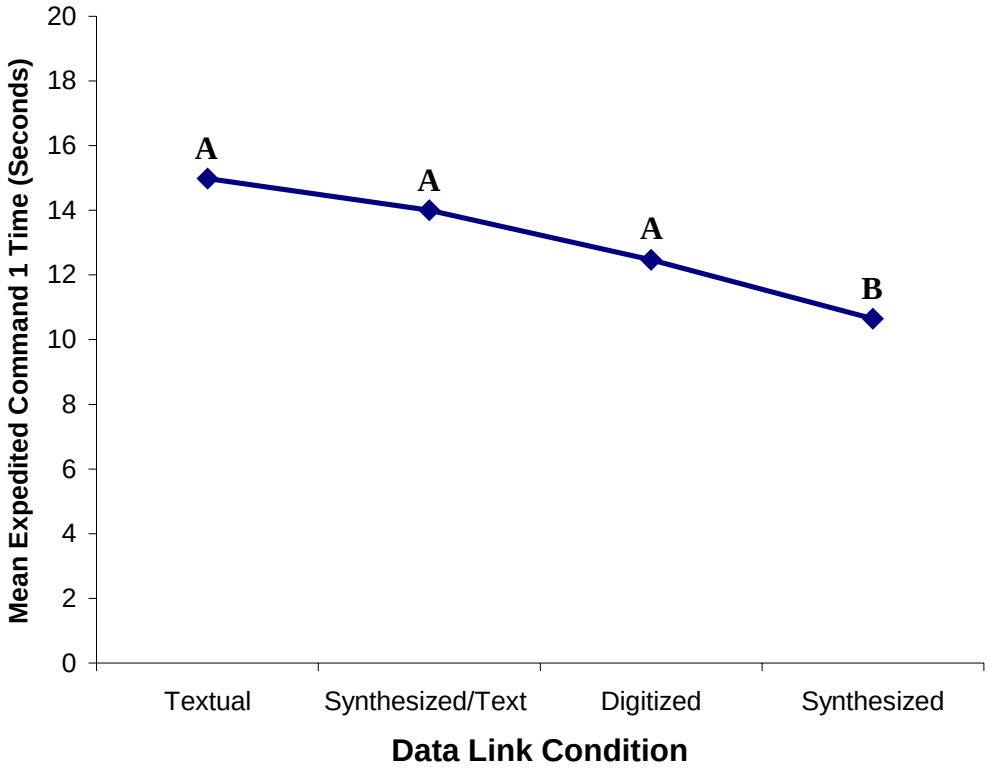


Figure 25. Mean expedited command 1 time for epoch 2 across data link conditions. Means with different letters are significantly different ($p < 0.05$) according to Tukey's test.

As was related previously, the improved intelligibility of the TTS that was used and the subjective impressions of its realism may be a factor. The key issue here is that, again, the textual data link condition resulted in increased expedited command 1 response time, which may have safety implications with respect to time-critical directives. As such, this represents further empirical data to suggest that textual data link presentations of emergency maneuvers are not desirable.

Expedited command 1 analysis for *head down time* revealed a significant main effect of data link condition, F (3,42) = 19.49, $p < 0.0001$, with the means ranging from 6.01 seconds to 9.93 seconds (see Table 19). Results are displayed in Figure 26. Post hoc analysis using Tukey's test revealed a significant difference between the textual data link and the other three data link conditions. The digitized data link resulted in significantly shorter time than did the textual data link condition, $t(42) = -5.98$, $p < 0.0001$. The synthesized data link condition also resulted in shorter time when compared to the textual data link condition, $t(42) = -6.79$, $p < 0.0001$. Finally, the synthesized/textual data link condition resulted in shorter time than did the textual data link condition, $t(42) = -5.78$, $p < 0.0001$.

The results for this metric are in line with what has been discovered thus far: that the textual data link condition represents increased response time as well as increased workload when compared to the other data link conditions. Thus, these additional empirical results support the contention that textual data link is undesirable for single pilot GA operations.

TABLE 19

Analysis of variance for expedited command 1, head down time

Source of Variance	df	MS	*F*	*p*
Between				
Flight Condition	1	15.97	3.32	0.0897
Subjects (S/FC)	14	4.81		
Within				
Data Link Condition (DL)	3	52.01	19.49	< 0.0001*
DL X FC	3	1.07	0.40	0.7522
DL X FC X S/FC	42	2.67		

*indicates significant result ($p \leq 0.05$)

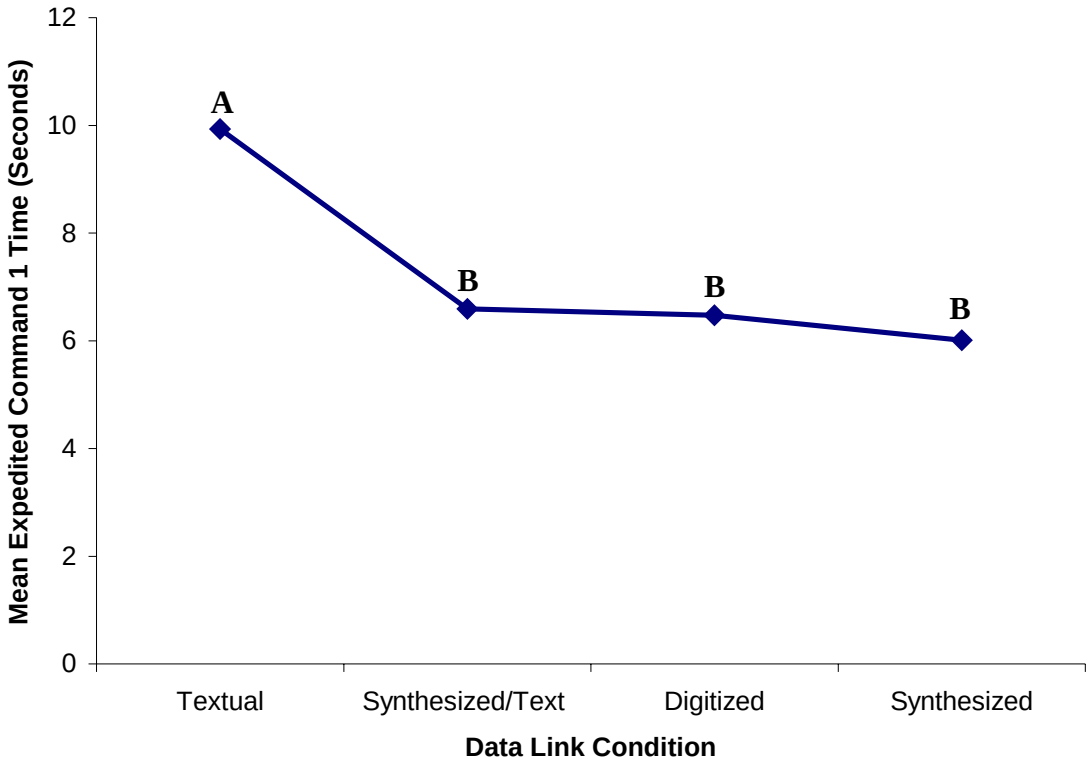


Figure 26. Mean expedited command 1 head down time. Means with different letters are significantly different ($p < 0.05$) according to Tukey's test.

Expedited command 2 analysis for head down time revealed a significant interaction between flight condition and data link condition, $F(3,42) = 8.04$, $p = 0.0002$, with the means ranging from 5.13 seconds to 7.97 seconds (see Table 20). Results are displayed in Figure 27.

Having a significant interaction within this data set means that the effects of the four data link conditions with respect to expedited command 2 head down time varied between the VFR and MVFR flight conditions. Therefore, data link conditional effects were not independent of flight condition effects. The most useful strategy to examine and interpret how the interaction means are affected is to again look at a 'slice' of the flight conditions to see how the data link conditions varied between VFR and MVFR.

TABLE 20

Analysis of variance for expedited command 2, head down time

Source of Variance	**df**	**MS**	***F***	***p***
Between				
Flight Condition (FC)	1	1.32	0.68	0.4227
Subjects (S/FC)	14	1.95		
Within				
Data Link Condition (DL)	3	6.40	1.88	0.1471
DL X FC	3	27.40	8.04	0.0002*
DL X FC X S/FC	42	3.40		

*indicates significant result ($p \leq 0.05$)

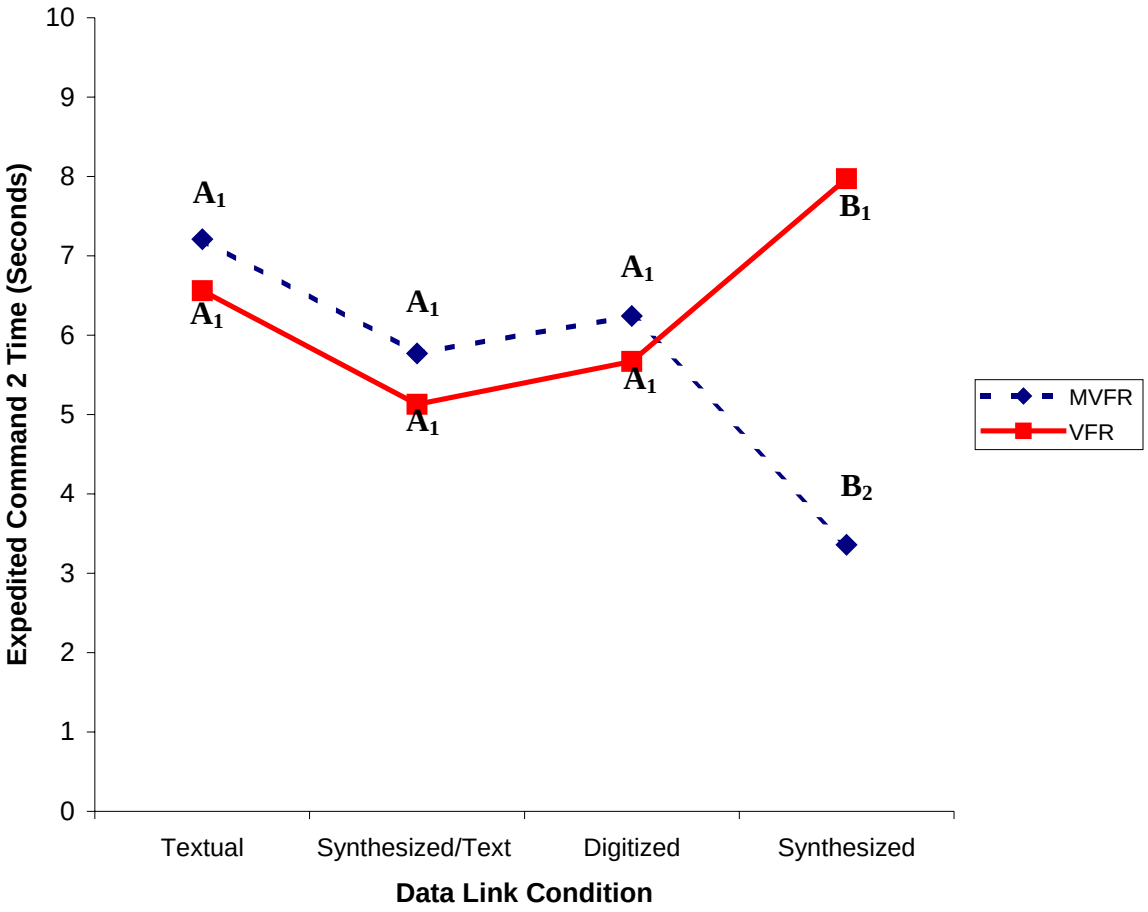


Figure 27. Interaction effect of expedited command 2 on head down time. Within a given flight condition, means with different letters are significantly different ($p < 0.05$) according to Tukey's test. Within a given data link condition, means with different numerals are significantly different ($p < 0.05$) according to Tukey's test.

Within the MVFR data set, Tukey's test revealed that the synthesized data link condition resulted in significantly less head down time than did the textual data link condition, $t(42) = -4.17$, $p = 0.003$. Within the VFR data set, however, the synthesized data link condition took significantly *more* head down time the did textual data link condition, $t(42) = -3.99$, $p = 0.005$.

The effect of the synthesized data link condition on expedited command 2 head down time is quite striking: there appears to be a slight advantage within the VFR flight

condition over the MVFR flight condition with respect to head down time but a significant difference in data link is present when synthesized speech alone is used. Post-hoc analysis using Tukey's test also reveals a significant difference between MVFR/synthesized and VFR/synthesized head down times, $t(42) = -3.99$, $p = 0.005$, with the MVFR condition resulting in shorter time. The synthesized data link condition results in a 4.6 second advantage within MVFR over VFR in expedited command 2 head down time. The reasoning behind this result may lie in the contention that the pilots perceived some level of urgency through the synthesized utterance that was not evident within the other data link conditions, similar to the results for epoch 1 time. The degraded visual conditions within MVFR may have provided some additional level of arousal such that, when coupled with synthesized speech, resulted in a performance boost for this measure. Why this effect is not mirrored within the synthesized/textual data link condition, which presents the same aural stimulus, is an open question.

That the synthesized speech utterances may have affected the pilots' perceived urgency (and thus arousal) has some support. With respect to the theory of information transfer developed by Shannon and Weaver (1949) and expanded by Finn (1977), 'typical' messages from ATC (defined here as 'frequent') may have less 'lexical complexity' (defined earlier) than 'atypical' messages (infrequent) and may thus lead to some level of increased arousal. Information is a function of two variables: 1) the number of lexical markers associated from the word, and 2) the number of lexical markers supplied from the word, or its 'transfer feature.' The theory states that frequent words are 'low information words' because they have few lexical markers. Conversely, rare words are 'high information words' because they have many markers that differ from

normal. This suggests that the infrequent, atypical expedited messages, in concert with the increased workload that MVFR rule maintenance requires, may have positively affected expedited command head down time when using synthesized voice. The Shannon and Weaver research was motivated by the desire to increase the *efficiency* and *accuracy* or fidelity of transmission and reception. 'Efficiency' refers to the bits of information per second that can be sent and received. 'Accuracy' is the extent to which signals of information can be understood. As such, accuracy refers more to clear reception than to the meaning of message. It may be that the synthesized utterances may have had an impact on both indices—the artificial voice presented some level of arousal such that the atypical expedited message had some positive impact on the efficiency and the accuracy of the message and thus overall data link performance.

These indexes can be further optimized through the use of speech production systems that can impart stress in an effort to increase efficiency as well as accuracy. Bou-Ghazale relates a method with which to identify indicators of stress and formulate novel statistical models to characterize speech parameter variation under stress (1997). The proposed models could then be integrated within speech synthesis and recognition systems to improve the naturalness of synthetic speech and recognition of speech under stress. That the models could also be applied to modify the speaking style of any new input speaker in such a way as to 'convince' pilots that the modified speech is under stress could further impact expedited response time.

Situation awareness

SAGAT. Fisher's Exact test is appropriate for this data set because the data are binomial (i.e., 1 = correct, 0 = incorrect) and thus represent nonparametric indexes. Fisher's test was used to analyze each of the twelve SAGAT queries. Fisher's Exact test indicates that SAGAT query 1 did not have a uniform SAGAT response across data link conditions ($p = 0.02$); see Table 21.

TABLE 21

Contingency table, data link conditions vs. incorrect or correct answer as measured using the SAGAT technique.

Frequency *Percent* *Row Percent* *Col Percent*	Incorrect Answer (0)	Correct Answer (1)	Total
Textual	1 1.56% 6.25% 9.09%	15 23.44% 93.75% 28.30%	16 25.00
Synthesized Speech & Textual	4 6.25% 25.00% 36.36%	12 18.75% 75.00% 22.64%	16 25.00
Digitized	0 0.00 0.00 0.00	16 25.00% 100.00% 30.19%	16 25.00
Synthesized Speech	6 9.38% 37.50% 54.55%	10 15.63% 62.50% 18.87%	16 25.00
Total	11 17.19	53 82.81	64 100.00

Figure 28 displays this non-uniformity of SAGAT query 1 response across data link conditions. What is immediately notable is that the digitized data link condition resulted in a 100% correct response rate across data link conditions. No other data link condition enjoyed a 100% correct response rate with respect to SAGAT query 1, although the textual data link came close with only one incorrect response. The synthesized/textual data link resulted in four correct answers. However, the synthesized data link condition resulted in six pilots (38%) reporting an incorrect response for SAGAT query 1. SAGAT query 1 was: "*What is your current deviation from your intended/assigned heading?*"

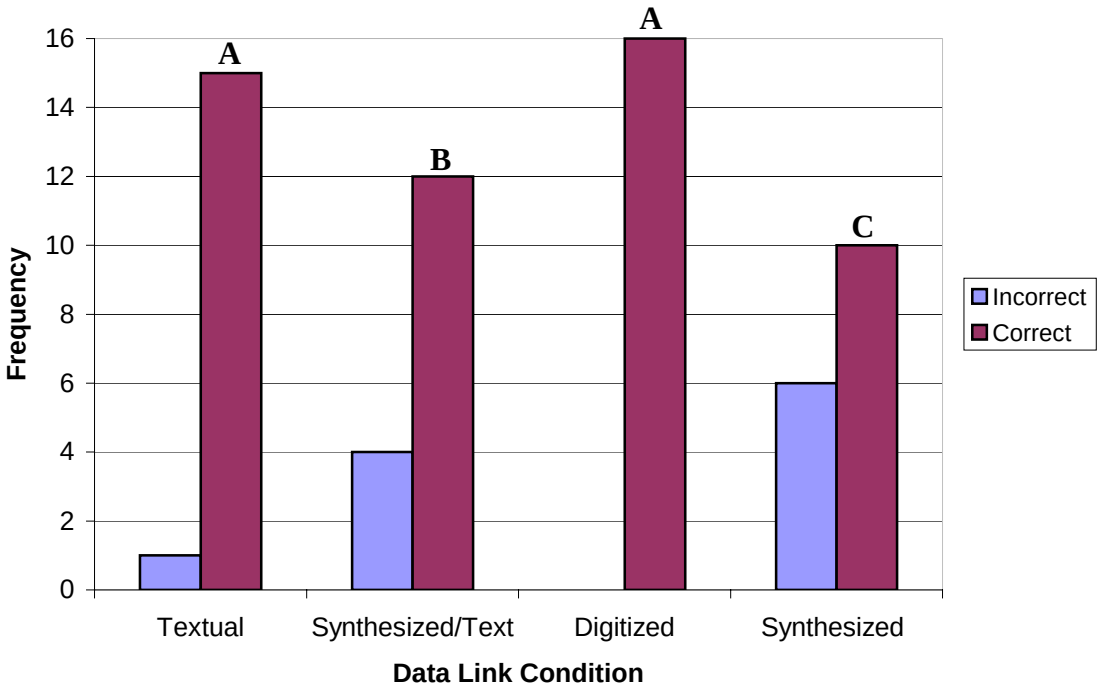


Figure 28. SAGAT query 1 response across data link conditions. Distributions with different letters are significantly different ($p < 0.05$) according to Fisher's test.

When rerunning Fisher's test to compare two conditions at a time, the synthesized data link resulted in more incorrect responses when compared with the other data link conditions ($p < 0.0001$). Fisher's test also revealed a slight advantage for the

synthesized/textual data link condition over the synthesized data link with respect to SAGAT response ($p < 0.05$). The digitized and textual data links did not differ with respect to SAGAT response performance.

Query 1 represented a 'Level 2' query, which investigated the pilots' ability to comprehend and integrate what is happening *right now* in the situation. As such, the difference in response rates within the synthesized data link condition suggests that the synthesized speech directives may have presented difficulties for the pilots in committing to their short term memory stores the index of heading. The complex situation in which the pilots found themselves, especially one in which a new and novel system is a part (i.e., data link), may have had an effect on the pilots' abilities to develop and maintain an adequate mental model with respect to their aircraft position and its deviation from where it should be based on synthesized speech directives. The fact that the only other data link condition resulting in two or more incorrect responses was synthesized/text provides a further indication that the synthesized data link presentation had some systematic effect on mental model construction and maintenance of the flight variables related to the recall of aircraft azimuth position. Recall that the digitized condition resulted in a 100% correct response rate. It may be that the recorded human speech within the digitized data link condition positively affected the pilots' SA when compared to the synthesized data link condition's *concatenated* human speech. That is, perhaps the natural flow of articulation (prosody) evident in the digitized speech directives fostered retention and recall of heading deviation over the (comparatively) artificial nature of the directives emanating from the TTS engine. It may be that advances in TTS output have not yet reached a level of maturity to be a useful alternative to human speech with respect to supporting SA.

SART. The SART responses were tested for normality as a justification for ANOVA. The analysis indicated that the responses represented normally distributed (i.e., parametric) data (see Table 23). ANOVA was then conducted on the SART responses.

TABLE 23

SART tests for normality

Test	Value	*p*
Shapiro-Wilk	0.98	0.47
Kolmogorov-Smirnov	0.08	0.15
Cramer-von Mises	0.06	0.25
Anderson-Darling	0.33	0.25

Analysis of pilot subjective SA ratings revealed a significant effect of pilots within flight condition, $F(1,14) = 15.04$, $p < 0.0001$, with the means ranging from 27.75 to 40.25 within VFR and from 21 to 58.5 within MVFR (see Table 24). Pilot SA ratings are displayed in Figure 29. Pilots varied across flight condition in their perceptions of their SA. What is immediately apparent in Figure 29 is that there exists more variation in perceived subjective SA within the MVFR conditions than in the VFR conditions. Within the VFR conditions, pilots reported an average SART rating of 33.7. Using Tukey's test, only one pilot within the VFR data set differed from the other pilots in subjective SA ($p < 0.002$). Within the MVFR condition, pilots reported an average SART rating of 33.1. Subject 14 reported the highest SA (58.5), which differed from the other pilots ($p < 0.0001$). Subject 10 reported the lowest SA (21), which also differed from the other pilots ($p < 0.0001$).

The large variation within the MVFR conditions with respect to reported SA is likely a result of a combination of two factors: 1) an *inflated sense of SA*, and 2) the

degraded environmental conditions within MVFR. With respect to the former, highly experienced pilots may maintain an inflated confidence in their own abilities, which is typical of commercial and military pilots. Anecdotally, it has been related that pilots who are 'intermediate' (i.e., operationally defined here as a low time pilot with total flight hours between 250-500) have an over-inflated sense of their own flight skills. Three subjects within the MVFR data set had flight hour totals in this range. Subjects 14 and 15, with total flight hours of 250 and 480, respectively, reported the highest SA; however, subject 10 (325 total flight hours), also falls into this range but reported lower SA, so this contention is open to debate.

TABLE 24

Analysis of variance for SART ratings

Source of Variance	**df**	**MS**	***F***	***p***
Flight Condition (FC)	1	5.06	0.24	0.4227
Subjects (S/FC)	14	321.35	15.04	<0.0001*
Data Link Condition (DL)	3	51.08	2.39	0.0822
DL X FC	3	30.06	1.41	0.2542
DL X FC X S/FC	42	21.37		

*indicates significant result ($p \leq 0.05$)

The second factor (degraded MVFR conditions) may have also played a role. The uniform level of perceived SA within the VFR data set may result from the added 'freedom of movement' within that flight condition which the MVFR pilots did not enjoy. The MVFR pilots were tasked with meeting the rules of cloud separation, which required constant maintenance of aircraft position with respect to distance from clouds.

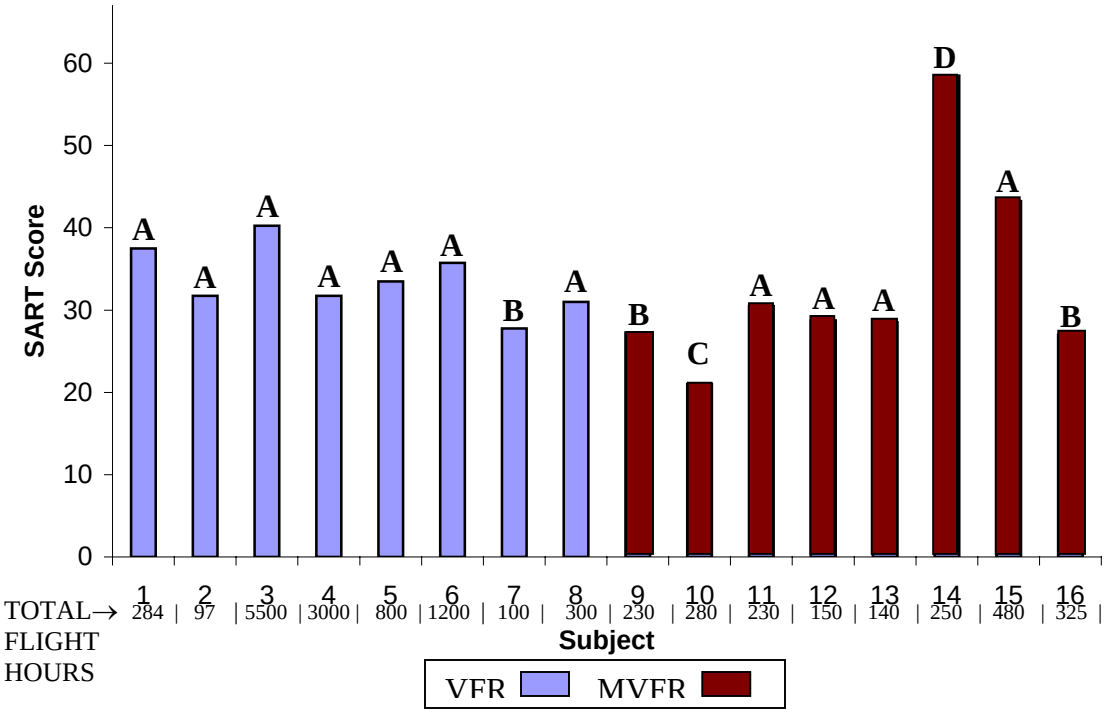


Figure 29. SART scores across flight conditions. Means with different letters are significantly different ($p < 0.05$) according to Tukey's test.

Pilots must maintain, in *any* condition, 500 feet below, 1000 feet above, and 2000 feet horizontal distance from clouds. Within MVFR, conditions are so degraded that there is very little room for deviation before the aforementioned rules are violated. Thus, workload begins to affect SA: the resources required to meet the demands of MVFR rule maintenance begin to interfere with construction and maintenance of the mental models that support SA. However, as the workload results did not indicate differences in either perceived or measured workload across flight conditions, this contention is not fully supported.

Questionnaire

The questionnaire can be viewed in Appendix F. Fisher's Exact test was appropriate for this data set because the data represented discrete categories, and conclusions about the respondents' distributions were of more interest than mean responses (e.g., did more pilots 'agree' with a particular question regarding data link presentation). As such, the only way to answer the questions presented was to analyze the frequency of responses as opposed to analyzing each individual pilot's response. Fisher's test was used to analyze each of the seven questionnaire items between flight conditions and across data link conditions.

TABLE 25

Contingency table, auditory data link conditions vs. voice articulation rating as measured via questionnaire.

Frequency *Percent* *Row Percent* *Col Percent*	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Digitized	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	11 22.92% 68.75% 44.00%	5 10.42% 31.25% 62.50%	16 33.33
Synthesized	0 0.00 0.00 0.00	2 4.17% 12.50% 40.00%	4 8.33% 25.00% 44.44%	9 18.75% 56.25% 36.00%	1 2.08% 6.25% 12.50%	16 33.33
Synthesized/Text	1 2.08% 6.25% 100.00%	3 6.25% 18.75% 60.00%	5 10.42% 31.25% 55.56%	5 10.42% 31.25% 20.00%	2 4.17% 12.50% 25.00%	16 33.33
Total	1 2.08	5 10.42	9 18.75	25 52.08	8 16.67	48 100.00

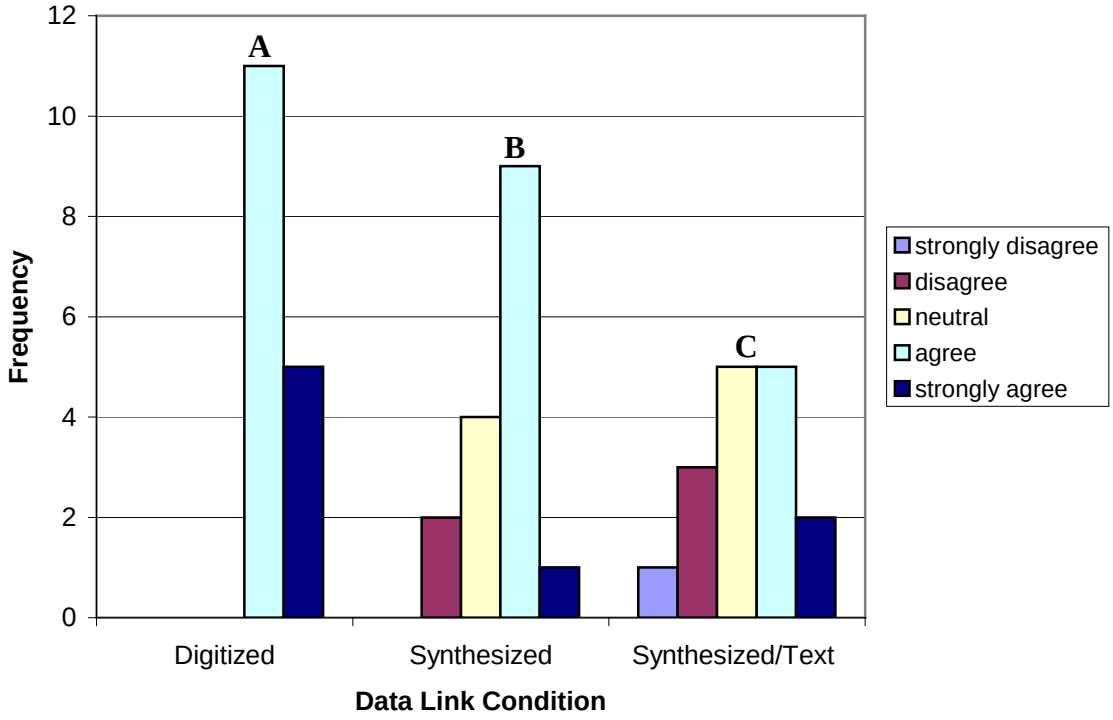


Figure 30. Questionnaire results for voice articulation. Distributions with different letters are significantly different ($p < 0.05$) according to Fisher's test.

Fisher's test indicated that subjects did not have a uniform perception of voice articulation across data link condition ($p = 0.02$). Results for articulation are displayed in Table 25. Figure 30 displays this non-uniformity of response for articulation across data link conditions. When rerunning Fisher's test to compare two conditions at a time, all three data link conditions were found to differ ($p < 0.05$). With respect to the digitized data link condition, all pilots agreed that the articulation within it was satisfactory. Differences in perceived articulation appear within the synthesized data link condition, and are even more variable within the synthesized/text data link condition. Ten pilots agreed that the articulation of the TTS engine was satisfactory, with four neutral and two disagreeing with this statement.

Interestingly, a majority of the pilots agreed that articulation was satisfactory within the synthesized/text data link condition, with four disagreeing and five neutral—even though the aural stimulus was exactly the same. Perhaps the inclusion of the textual element within this data link condition negatively affected pilots' perceptions of the synthesized voice, for whatever reason. Subjective comments were mixed with respect to the digitized data link condition: *"(It was) easier to understand both common and uncommon words than with (the) generated voice"* while others were not as supportive: *"the voice was acceptable, but not as clear as synthesized."* With respect to the synthesized data link conditions, pilots were also mixed in their impressions of the voice: *"pronunciation (was) a little scratchy"* and *"It was a lot slower and easier to understand than most real controllers. For the most part, everything was clear, but it sounded too robotic"; "clear, understandable. Sounds synthesized but not unpleasant."* Thus, with respect to the articulation of data link commands, the digitized voice appears to be more desirable than synthesized voice. Perhaps the strides that have been made in recent years in speech synthesis technology and thus perceived TTS engine articulation quality still have not met the expected qualities of human voice, as suggested previously. As a result, such systems may not be indicated for aural ATC directives with respect to data link until they reach a more mature level.

Fisher's Exact test also indicated that subjects did not have a uniform perception of voice naturalness across data link condition ($p < 0.00004$). Results for naturalness are displayed in Table 26. Figure 31 displays this non-uniformity of response for naturalness across data link conditions. When rerunning Fisher's test to compare two conditions at a time, all three data link conditions were again found to differ ($p < 0.05$). As with the

impressions of voice articulation, pilots indicated general agreement (one pilot was neutral) as to the naturalness of the digitized data link condition. As digitized voice is recorded human speech, this result is not surprising. Also as in the responses for voice articulation, pilots were mixed in their impressions of voice naturalness, with more variation within the synthesized/text data link condition than in the other two voice presentations. Digitized speech enjoyed almost universal agreement as to its naturalness (one pilot was neutral). Subjective comments support this contention: *"I prefer natural speech over the artificial one"* and *"very natural speech—close to day to day voice from ATC. I felt I was more accurate with altitudes and headings than in the other scenarios."* What is most notable is that the same number of pilots generally disagreed **and** generally agreed that the synthesized voice was natural sounding in both data link conditions that incorporated synthesized speech.

The number of pilots who reported strong disagreement with that statement for synthesized speech increased four-fold when evaluating the synthesized/text combination. Again, these results are hard to discern as both synthesized data link conditions incorporated that same aural stimulus. As such, interpretation of the pilots' perceptions as to the naturalness of synthesized speech is difficult. However, the results do follow the same general trend as was evidenced in the voice articulation responses: that the TTS engine technology used in the current research (i.e., AT&T's Natural Voices) and the quality of its output may not have reached a level of maturity to be acceptable to pilots for use in the cockpit.

TABLE 26

Contingency table, auditory data link conditions vs. voice naturalness rating as measured via questionnaire.

Frequency *Percent* *Row Percent* *Col Percent*	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Digitized	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 2.08% 6.25% 14.29%	9 18.75% 56.25% 64.29%	6 12.50% 37.50% 85.71%	16 33.33
Synthesized	1 2.08% 6.25% 20.00%	9 18.75% 56.25% 60.00%	3 6.25% 18.75% 42.86%	3 6.25% 18.75% 21.43%	0 0.00 0.00 0.00	16 33.33
Synthesized/Text	4 8.33% 25.00% 80.00%	6 12.50% 37.50% 40.00%	3 6.25% 18.75% 42.86%	2 4.17% 12.50% 14.29%	1 2.08% 6.25% 14.29%	16 33.33
Total	1 10.42	15 31.25	7 14.58	14 29.17	7 14.58	48 100.00

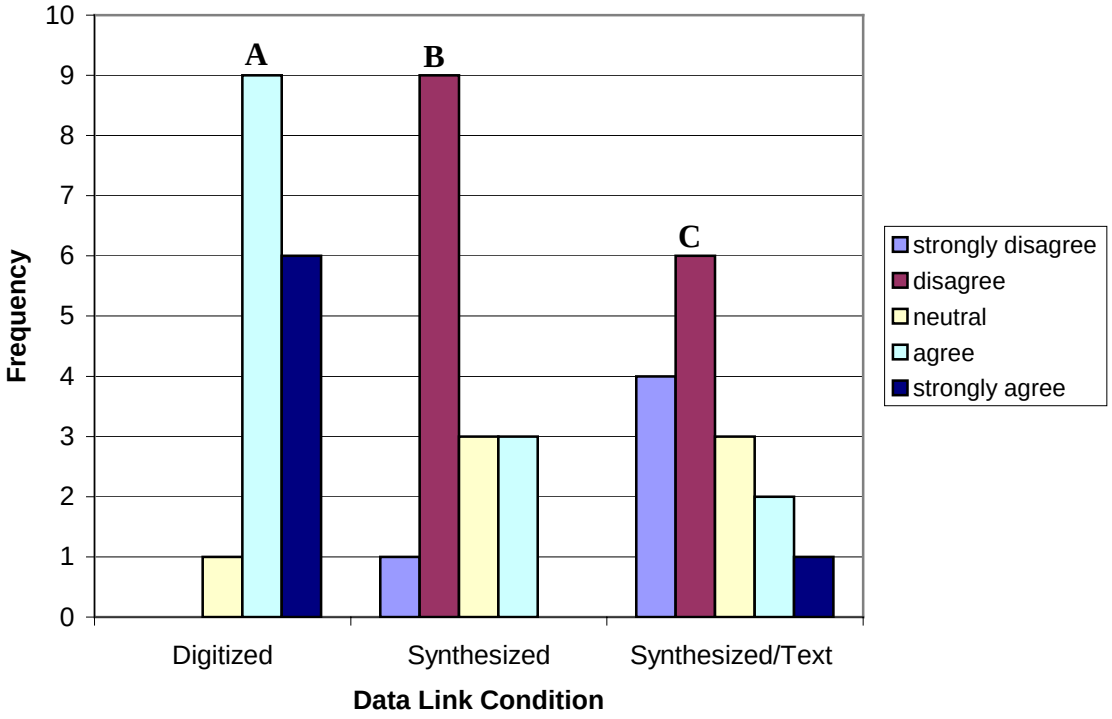


Figure 31. Questionnaire results for voice naturalness. Distributions with different letters are significantly different ($p < 0.05$) according to Fisher's test.

Fisher's Exact test also indicated that subjects did not have a uniform perception of voice stress across data link condition ($p < 0.00004$). Results for voice stress are displayed in Table 27. Figure 32 displays this non-uniformity of response for stress across data link conditions. When rerunning Fisher's test to compare two conditions at a time, all three data link conditions were found to differ ($p < 0.05$). No pilots disagreed strongly with the statement that stress was applied appropriately on words that required it, although many pilots expressed disagreement that either synthesized data link condition appropriately applied stress. Within the synthesized data link condition, the same number of pilots agreed as disagreed that the utterances presented with appropriate stress. The synthesized/text data link condition displays what appears to be a negative trend toward disagreement, with more pilots disagreeing or reporting a neutral stance with respect to voice stress than indicated agreement.

As was related previously, every effort was made when recording the digitized directives to ensure that stress was *not* applied because of the inability of the TTS engine to do so. The fact that no pilots disagreed with the digitized data link condition's appropriate use of stress indicates that these efforts were not as successful as was intended—it appears that some element of stress was recorded anyway. The questionnaire also maintained an open-ended section wherein pilots could relate their impressions of the textual data link.

Many pilots expressed severe reservations in the adoption and use of a text-only data link: "*it (was) increasingly difficult to read a screen as well as transfer the information to the flight controls.*

TABLE 27

Contingency table, auditory data link conditions vs. voice stress rating as measured via questionnaire.

Frequency *Percent* *Row Percent* *Col Percent*	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Digitized	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 8.33% 25.00% 33.33%	9 18.75% 56.25% 50.00%	3 6.25% 18.75% 60.00%	16 33.33
Synthesized	0 0.00 0.00 0.00	6 12.50% 37.50% 46.15%	3 6.25% 18.75% 25.00%	6 12.50% 37.50% 33.33%	1 2.08% 6.25% 20.00%	16 33.33
Synthesized/Text	0 0.00 0.00 0.00	7 14.58% 43.75% 53.85%	5 10.42% 31.25% 41.67%	3 6.25% 18.75% 16.67%	1 2.08% 6.25% 20.00%	16 33.33
Total	0 0.00	13 27.08	12 25.00	14 29.17	7 14.58	48 100.00

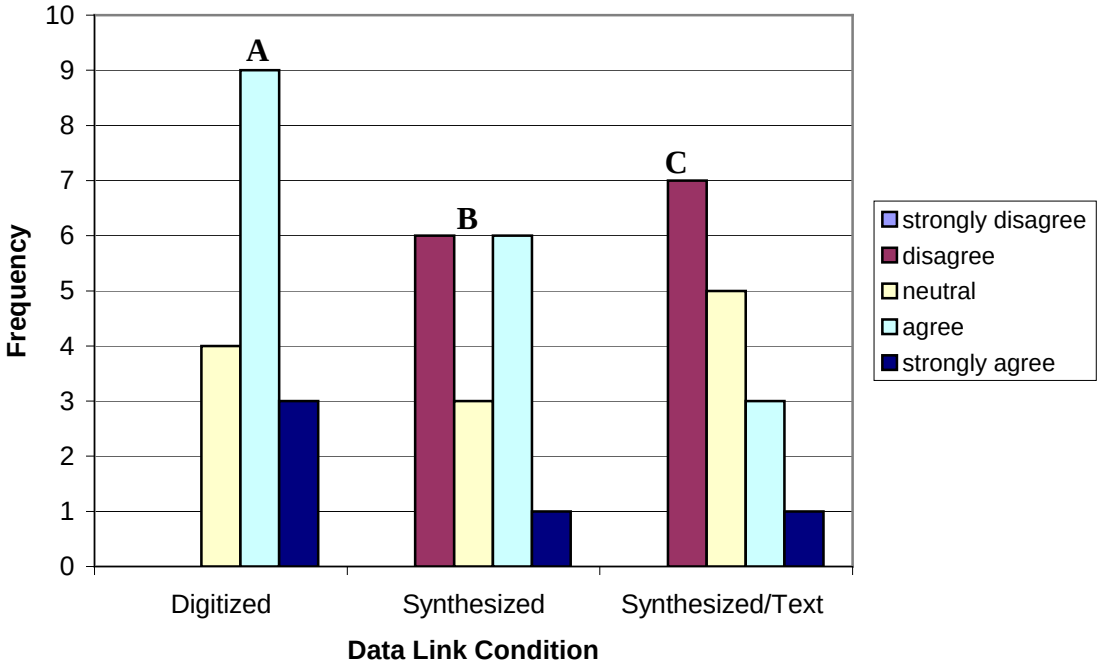


Figure 32. Questionnaire results for voice stress. Distributions with different letters are significantly different ($p < 0.05$) according to Fisher's test.

I felt it took too much time to read the commands, comprehend them, (and) then respond to them. This increased workload and decreased the efficiency of actually flying the aircraft" and *"(It was) difficult to follow sheer written instruction, (it) keeps you inside the cockpit. (It) takes too long for emergency instructions."* These concerns are evidenced in the workload results described previously. Many pilots thought that the synthesized/text data link condition was the most useful: *"(It was) much easier because text data helped to confirm voice data"* and *"(It was) good to have the text backup in case I misunderstand or it is tough to understand, but not HAVING to read is great."*

CONCLUSIONS

The workload results in this study, which were both subjectively and objectively determined, supported each other quite nicely. That is, the subjective MCH results, which indicated unacceptable levels of workload within the textual data link condition, were mirrored in the objective head down time analysis, which indicated increased head down time within the textual data link condition. The data suggest that a textual data link presentation *alone*, or without some aural component, is not indicated for single-pilot GA operations. It was interesting to note that pilot experience (as evidenced in reported total flight hours) did not appear to have an effect on perceived workload, especially within the textual data link condition. Further, pilots did not report workload ratings that differed between flight conditions. This was not expected as rule maintenance (i.e., cloud separation) within MVFR conditions should have been more difficult to accomplish due to the severely degraded outside view. However, the fact that no difference was found between flight condition suggests that this is an issue of simulator fidelity—the simulator's outside view, which was relegated to a 1.5-inch section of the flat panel screen, simply may not have been able to adequately render outside scenes such that the required visual cues with which to satisfy MVFR rules were evident. An effort was made to locate published research and/or gather anecdotal information concerning the threshold level of speech intelligibility that is necessary for pilots to perform in high workload situations, but this effort was unsuccessful. Such data would have been useful in exploring the relationship between speech intelligibility and workload, which has implications for auditory displays in the cockpit.

Epoch analyses resulted in several interesting findings. The interaction effect of epoch 1 (the time from alert until 'receive message') was surprising. As the alerting stimuli (i.e., bell chime and blinking light) did not differ across conditions, differences in epoch 1 time would not have been expected. But the fact that the data link conditions incorporating synthesized speech resulted in shorter epoch 1 times than both the digitized and textual data link conditions (within the MVFR data set) as well as from text (in the VFR data set) may suggest that some element of 'expectation' for synthesized speech exists. The epoch 2 analyses, which represented the main focus of this study (i.e., the time between accessing and responding to the message, which is dependent on the modality of data link), resulted in quite a compelling case against the use of a textual-only data link system. The textual data link clearly took longer than the other three data link conditions in epoch 2 time. The pilots were required to read the data link message, which existed on the data link touch screen, and the time required to complete that task was significantly increased over the time required to hear that same message aurally. These results are supported by the workload results, which indicated a perceived unsafe condition when using textual data link. As most research into data link has focused solely on pilot teams, integration of a textual-only data link into GA, which is comprised largely of single pilot operations, is a dangerous avenue to take.

Expedited command analysis was valuable in that single pilot data link integration could be evaluated with a focus on time-critical or emergency responses. As before, the textual data link condition resulted in the longest epoch 2 time, providing further evidence for the contraindication of this modality, with similar significance for reduced expedited command performance as determined through head down time analysis.

Synthesized data link was found to be significantly decreased for expedited command time, which may indicate that the intelligibility of the TTS engine that was used was sufficiently realistic to effect performance. This contention is further explored in the interaction effect in expedited performance for head down time, which suggested some level of perceived urgency for synthesized voice that was not evident within the other data link conditions. However, these results must be tempered with the fact that similar results were not obtained for the synthesized/textual data link, which maintained the same aural stimulus, as did the synthesized data link. While this research did not attempt to provide definitive answers to questions regarding differences in performance between 'normal' (i.e., heading, altitude change commands) and 'emergency' (i.e., time-critical commands) data link messages, it did begin to explore the issue of presentation format for emergency directives, and helps point to future research in that regard, such as the development and testing of TTS systems that can incorporate stress elements that may foster performance improvements.

For situation awareness, which was measured objectively and subjectively (i.e., SAGAT and SART, respectively), the results are mixed. Of the 12 SAGAT queries, only one was found to differ across data link conditions—that of azimuth position recall. This query was found to be incorrect more often within the synthesized data link condition than in the other conditions, and represents a 'Level 2' query, which investigates the pilots' ability to comprehend and integrate what is happening *right now* in the situation. As such, the difference in response rates within the synthesized data link condition suggests that the synthesized speech directives may have presented difficulties for the pilots in committing to their short term memory stores the index of heading. The

complex situation in which the pilots found themselves, especially one in which a new and novel system is a part (i.e., data link), may have had an effect on the pilots' abilities to develop and maintain an adequate mental model with respect to their aircraft position and its deviation from where it should be based on synthesized speech directives. Thus, it may be that advances in TTS output have not yet reached a level of maturity to be a useful alternative to human speech with respect to supporting SA. The SART analysis indicated that the pilots varied across flight condition in their perception of their SA—there existed more variation in perceived subjective SA within the MVFR conditions than in the VFR conditions. This variation within the MVFR conditions with respect to reported SA is likely a result of a combination of two factors: 1) an *inflated sense of SA*, and 2) the degraded environmental conditions within MVFR. With respect to the former, highly experienced pilots may maintain an inflated confidence in their own abilities, which is typical of commercial and military pilots. Anecdotally, it has been related that pilots who are 'intermediate' (i.e., operationally defined as a low time pilot with total flight hours between 250-500) have an over-inflated sense of their own flight skills. Within MVFR, conditions are so degraded that there is very little room for deviation before the aforementioned cloud separation rules are violated. Thus, workload begins to affect SA: the resources required to meet the demands of MVFR rule maintenance begin to interfere with construction and maintenance of the mental models that support SA. However, due to the workload results, which did not indicate differences in either perceived or measured workload across flight conditions, this contention is not fully supported.

The questionnaire results indicated differences in perceived voice articulation, naturalness, and stress across data link conditions. Pilots largely agreed that the digitized voice was the most natural, articulate, and provided adequate stress. As noted earlier, an effort was made to record the digitized utterances *without* stress, as the TTS engine utilized did not have the ability to do so. The result for stress suggests that this effort was not entirely successful. However, most pilots were supportive of the TTS engine's ability to articulate data link commands and sound natural while doing it. This is an important finding as it suggests that improvements in technology are resulting in improved realism, which may have implications for the integration and acceptance of auditory displays in the cockpit that utilize synthesized voice. Most pilots were very supportive of the redundancy evident in the synthesized/textual data link presentation, which is not too surprising in that redundant cues would be expected to provide support in systems that maintain a single operator.

The results of the research described herein provide data to the designers of future aviation regimes specific to single-pilot GA operations that attempt incorporate increased automation in the form of data link systems. Single pilot GA operations do not enjoy the presence of another human in the cockpit who can support and verify the activities of flight, as do the pilot teams endemic in commercial and military operations. Again, what is quite clear from the analyses presented is the detrimental affect of a text-only data link presentation on all measured aspects of the experimental trials (i.e., performance, workload, and situation awareness). Text-only data link, which is the current flavor of CPDLS within commercial and military operations, is not indicated for implementation into single-pilot GA operations. Redundant cues (e.g., the two data link alerting

schemes) are an absolute necessity in dynamic systems in which there is a single operator. Such redundancy can foster a desirable state in the single GA pilot such that he/she operates in an atmosphere that supports and maintains low levels of workload and improved situation awareness. Similarly, the use of redundancy in the form of a textual backup of aural data link directives is suggested for the integration of data link into GA. With respect to the aural component of any proposed data link system, the data appear to suggest that there are no differences between a digitized or synthesized presentation from a performance standpoint, although some variation exists with respect to subjective preference (related to the refined yet apparently still immature technology of the TTS system used). Performance between the three aural data link presentations did not differ, but at least one analysis (expedited command performance) indicated a slight temporal advantage for the synthesized speech data link condition.

FUTURE RESEARCH

As discussed previously, early data link investigations using commercial flight teams found measurable performance improvements when using synthesized voice 'callouts' (e.g., altitude, airspeed) over those presented from the pilot-not-flying. Given the technological improvements as of late, especially that of the global positioning system, altitude radar, and infrared systems, it appears there may be benefit to investigating such synthesized call-outs within single-pilot GA operations. Much like the dependent measures utilized by Simpson (1981), wherein flight performance was a measure of the percentage of time 'out of operational tolerance for flight parameters,' similar measures could be utilized within the GA environment (i.e., percentage of time spent 'out of the glide slope'). Another consideration is the use of voice technology for direction prompting or way-finding tasks, similar to how such displays are used currently in automobiles (e.g., Liu, 2001; Liu, Schreiner, and Dingus, 2000). Pilots could conceivably use voice prompting, for example, to navigate via very-high frequency omindirectional range (VOR), which already has an auditory component in the form of an alerting tone. Changing those tones to a voice prompt that could relate progress (in nm) to or from a particular VOR might be beneficial in further reducing the head-down time required in monitoring the VOR needle. Other possible applications include future tower control systems wherein aspects of control may be automated. An example of this could be seen in ground control operations, where automated technologies are already in place at some airports (e.g., infrared identification of aircraft taxi position within heavy fog conditions) and can conceivably use automated voice prompting (e.g., "proceed with caution to runway 33; be advised United 98 heavy will be ahead of you"). Other

applications of voice technology in future cockpits include callouts of system information (e.g., “manifold temperature above normal”) or as supplements to HITS displays (e.g., “enter O’Hare runway 11 approach brackets in 5nm”).

The experimental outcome has also served to concentrate further investigations of a similar nature. The use of ANR headsets within GA operations appears to be suitable for the purposes of intelligibility due to the prevalence of aircraft noise that exists largely within the low frequencies wherein ANR is most effective. Although not specifically investigated in the current research, the intelligibility of synthesized speech in aircraft engine noise might be improved using ANR headsets when compared to traditional passive devices, and warrants further investigation. Additionally, the use of ANR headsets may have implications for workload or SA when compared to traditionally passive aviation headsets. That is, the noise-reduction capabilities of ANR headsets may result in less effort from the pilot in the interpretation of radio messages, which may in turn result in decreased operational workload and improved SA. While the current research did not delve into the vagaries of data link screen design or placement from a usability standpoint, suggestions can be made with respect to the independent variables that warrant further investigation. For example, the capacity to retrieve and review messages previously received and sent, different alerting schemes for routine and emergency or time-critical procedures, and front, rear, or central positioning of the data link interface are just a few.

REFERENCES

Air Transport Association (1991). Human factors requirements for data link, condensed version. ATA information transfer subcommittee, In Rehmann, A.J., 1997, Human factors recommendations for controller-pilot data link (CPDL) communication systems: a synthesis of research results and literature (Tech. Report ACT-350; DOT/FAA/CT-TN97/6) William J. Hughes Technical Center, New Jersey.

Amar, M.J., Hansman, R.J., Hannon, D.J., Vaneck, T.W., and Ghaudhry, A.I. (1995). Human subject evaluation of airport surface situational awareness using prototypical flight deck electronic taxi chart displays. (Tech. Report). Cambridge, MA: John A. Volpe National Transportation Systems Center, November.

Andre, A.D., Wickens, C.D., Moorman, L., and Boschelli, M.M. (1991). Display formatting techniques for improving situation awareness in the aircraft cockpit. The International Journal of Aviation Psychology, 1, 3, pp. 205-218.

Baldwin, C.L., Struckman-Johnson, D. (2002). Impact of speech presentation level on cognitive task performance: implications for auditory display design. Ergonomics, 45, 1, pp. 61-74.

Baldwin, C.L. and Schieber, F. (1995). Dual task assessment of age differences in mental workload with implications for driving. In *Proceedings of the 39th Annual Meeting of Human Factors and Ergonomics Society*, p. 167-172.

Ballas, J.A. (2000). The niche hypothesis: implications for auditory display design. In *Proceedings of the IEA 2000/HFES 2000 Congress*, pp. 3-718-3-721.

Battiste, V. and Johnson, N.H. (2002). An operation evaluation of ADS-B and CDTI during airport surface and final approach operations. In *Proceedings of the 46th Annual Human Factors and Ergonomics Society Meeting*, (pp. 36-41). Santa Monica, CA: Human Factors and Ergonomics Society.

Bauschat, J-M. (2001). On the Dependence of Pilot Task and Pilot Workload, In *AIAA Modeling and Simulation Technologies Conference (AIAA 2001-4189)*, pp.214-22.

Begault, D.R. (1993). Head-Up Auditory Displays for Traffic Collision Avoidance System Advisories: A Preliminary Investigation, Human Factors 35, 4, pp. 707-717.

Begault, D.R. (1998). Virtual Acoustics, Aeronautics, and Communications, Journal of the Audio Engineering Society, 46, 6, pp. 520-30.

Begault, D.R. and Pittman, M.T. (1996). Three-Dimensional Audio Versus Head-Down Traffic Alert and Collision Avoidance System Displays. The International Journal of Aviation Psychology, 6, 1, 79-93.

Begault, D.R. and Wenzel, E.M. (1992). Techniques and Applications for Binaural Sound Manipulation in Human-Machine Interfaces. The International Journal of Aviation Psychology, 2, 1, pp. 1-22.

Bell Laboratories. (2002). Bell Laboratories Projects. Retrieved June 14, 2002, from: http://www.bell-labs.com/org/1133/Research/

Bellenkes, A.H. (1999). Tactical Scan Breakdown and Loss of Situation Awareness. In *Proceedings of the 4th Annual Symposium on Situation Awareness*, p. 135.

Bellenkes, A.H., Wickens, C.D., Kramer, A.F. (1997). Visual Scanning and Pilot Expertise: The Role of Attentional Flexibility and Mental Model Development. Aviation, Space, and Environmental Medicine, 68, 7, pp.569-79.

Benoit, C., Grice, M. and Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. Speech Communication, 18, p.381-392.

Blom, H.A.P., Stroeve, S., Daams, J. and Nijhuis, H.B. (2001). Human Cognition Performance Model Based Evaluation of Air Traffic Safety. National Aerospace Laboratory NLR, pp. 1-10.

Blom, H.A.P. (1992). The Layered Safety Concept, An Integrated Approach to the Design and Validation of Air Traffic Management Enhancements. National Aerospace Laboratory NLR (NLR TP 92046 U).

Bou-Ghazale, S.E. (1997). Analysis, modeling and perturbation of speech under stress with applications to speech synthesis and recognition. Dissertation Abstracts International: Section B: the Sciences & Engineering 57(12-B), Jun 1997, 7648, US: Univ Microfilms International.

Casali, J.G. and Wierwille, W.W. (1983). A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load. Human Factors, 25, pp. 623-641.

Cotton, J.C. and McCauley, M.E. (1983). Voice technology design guides for Navy training systems: Final report for the period 23 April, 1980-2 January, 1982). Tech Report No. NAVTRAEQUIPCEN 80-C-0057-1). Orlando, FL: Naval Training Equipment Center.

Cowley, C.K. and Jones, D.M. (1992). Synthesized or Digitized? A Guide to the Use of Computer Speech. Applied Ergonomics, 23, 3, p. 172-176.

Deatherage, B.H. (1972). Auditory and other sensory forms of information presentation. In H.P Van Cott and R.G. Kinkade (Eds.), Human Engineering Guide to Equipment Design. Washington, DC: US Government Printing Office.

Delgou, C., Conte, S. and Sementina, C. (1998). Cognitive factors in the evaluation of synthetic speech. Speech Communication, 24, p. 153-168.

Dreyfus, S.E. (1981). Formal models vs. human situational understanding: inherent limitations on the modeling of expertise. Berkeley, CA: Operations Research Center, University of California.

Durso, F.T., Hackworth, C.A., Truitt, T.R., Crutchfield, J., Nikolic, D., and Manning, C.A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. Air Traffic Control Quarterly, 6, 1, p. 1-20.

Edworthy, J. (1994). The design and implementation of non-verbal auditory warnings. Applied Ergonomics, 25, p. 202-210.

Endsley, M. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors and Ergonomics Society 32nd Annual Meeting*, (pp. 97-101). Santa Monica, CA: Human Factors and Ergonomics Society.

Endsley, M. (1988b). Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE National Aerospace and Electronics Conference (NAECON)*, p. 789-795.

Endsley, M.R. (1993). Situation awareness in dynamic human decision making: measurement. In *Proceedings of the 1st International Conference on Situational Awareness in Complex Systems*, Orlando, February.

Endsley, M. (1995). Measurement of situation awareness in dynamic systems. Human Factors, 37, 1, pp. 65-84.

Endsley, M. (1996). Automation and Situation Awareness. In Automation and Human Performance: Theory and Applications, (pp. 163-81). Lawrence Erlbaum Associates, N.J.

Endsley, M.R. (1997, September). Situation awareness: the future of aviation systems. Presented at the Saab 60th Anniversary Symposium, Linkoping, Sweden.

Endsley, M.R. (1998). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors and Ergonomics Society 32nd Annual Meeting*, (pp. 97-101). Santa Monica, CA: Human Factors and Ergonomics Society.

Endsley, M.R. (1999). Situation awareness and human error: designing to support human performance. In *Proceedings of the High Consequences Systems Surety Conference,* (pp. 1-9).

Endsley, M. and Garland, D.J. (Eds.)(2000). Situation Awareness Analysis and Measurement. Mahwah, NJ: Lawrence Erlbaum and Associates.

Endsley, M.R., Sollenberger, R., and Stein, E. (2000). Situation awareness: a comparison of measures. In *Proceedings of the Human Performance, Situation Awareness and Automation: User Centered Design for the New Millennium Conference* (pp. 1-6).

Federal Aviation Administration. (1991). The National Plan for Aviation Human Factors. (draft). Washington, DC: Author.

Federal Aviation Administration (1995). National Plan for Civil Aviation Human Factors: An Initiative for Research and Application. Washington, DC: Author.

Finn, P. J. (1977). Word frequency, information theory, and cloze performance: A transfer feature theory of processing in reading. Reading Research Quarterly, 13, 4,1977; 1978, 508-537. International Reading Association, US.

Francis, A.L. and Nusbaum, H.C. (1999). The effect of lexical complexity on intelligibility. International Journal of Speech Technology, 3 pp.15-25.

Garner, K.T. and Assenmacher, T.J. (1996). In Gawron (Ed.), 2002, Guide for Selecting Workload and Situational Awareness Measures Workshop. Human Factors and Ergonomics Society 46th Annual Meeting, Baltimore, MD.

Gawron, V.J. (2002). Guide for Selecting Workload and Situational Awareness Measures Workshop. Human Factors and Ergonomics Society 46th Annual Meeting, Baltimore, MD.

Gawron, V.J., Weingarten, N.C., Hughes, T. and Adams, S. (1999). Verifying situational awareness associated with flight symbology. In *Proceedings of the 37th Aerospace Sciences Meeting and Exhibit* (AIAA paper 99-1093).

Gerratt, B.R. and Kreiman, J. (2001). Measuring vocal quality with speech synthesis. Journal of the Acoustical Society of America 110, 5, 2560-2566.

Gopher, D., and Donchin, E. (1986). Workload—An Examination of the Concept. In Handbook of Perception and Human Performance: Volume II. Cognitive Processes and Performance, Chapter 41. New York: Wiley.

Greene, B.G., Logan, J.S. and Pisoni, D.B. (1986). Perception of synthetic speech produced automatically by rule: intelligibility of eight text-to-speech systems. Behavior Research Methods, Instruments, and Computers 18, 2, p. 100-107.

Haas, E.C.M. and Casali, J.G. (1995). Perceived urgency of response time to multi-tone and frequency-modulated warning signals in broadband noise. Ergonomics, 38, 2313-2326.

Hagar, D.R. and Payne, D.G. (1996). Dual-task performance and visual attention switching. In *Proceedings of the40th Annual Meeting of the Human Factors and Ergonomics Society*, pp. 546-561.

Hancock, P.A., Meshkati, N., Robertson, M.M. (1985). Physiological Reflections of Mental Workload. Aviation, Space, and Environmental Medicine, 56, pp. 1110-1114.

Hale, S. and Piccione, D. (1992). Application of the Subjective Workload Assessment Technique to Aviation Test and Evaluation. In *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting*, pp. 1185-1189.

Hameluck, D.E. (1990). Mental Models, Mental Workload, and Instrument Scanning in Flight. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting*, pp. 76-80.

Hameluck, D.E. (1991). Relationship Between Mental Models and Scanning Behavior During Instrument Approaches. In *Proceedings of the 6th International Symposium on Aviation Psychology*, pp. 939-44.

Harvey, C.M., Reynolds, M., Pacley, A.L. Koubek, R.J., and Rehmann, A.J. (2002). Effects of the controller-to-pilot data link (datalink) on crew communication. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, Santa Monica, CA: Human Factors and Ergonomics Society.

Higgins, T.J. and Chignell, M.H. (1987). Cognitive Processes During Instrument Landing. In *Proceedings of the Human Factors Society 31st Annual Meeting*, pp. 1216-1220.

Hopkin, V.D. (1995). Human Factors in Air Traffic Control, Taylor and Francis, London, UK.

House, A.S., Williams, C.E., Hecker, M.H., and Kryter, K.D. (1965). Articulation testing methods: Consonantal differentiation with a closed-response set. Journal of the Acoustical Society of America, 37, 158-166.

Illman, P.E. (1995). The Pilot's Handbook of Aeronautical Knowledge, McGraw-Hill, New York, NY.

Jones, D.G. and Endsley, M.R. (1996). Sources of situation awareness errors in aviation. Aviation, Space and Environmental Medicine, 67, 6, pp. 507-512.

Jones, D.G. and Endsley, M.R. (2000). Can real-time probes provide a valid measure of situation awareness? In *Proceedings of the Human Performance, Situation Awareness, and Automation User-Centered Design for the New Millennium Conference*, p. 1-6.

Kantowitz, B.H. and Knight, J.L. (1976a). Testing tapping timesharing, II: auditory secondary task. Acta Psychologica—International Journal of Psychonomics, 40, pp. 343-362.

Kantowitz, B.H. 7 Knight, J.L. (1976b). On experimenter-limited processes. Psychological Review, 83, 6, pp. 502-507.

Kantowitz, B.H. and Knight, J.L. (1977). Inferring decay in short-term memory: the issue of capacity. Memory & Cognition, 5, 2, pp. 167-176.

Kryter, K.D. (1985). The Effects of Noise on Man. Orlando, FL, Academic Press, p. 61.

Labiale, G. (1990). In-car road information: comparisons of auditory and visual presentations. In *Proceedings of the Human Factors and Ergonomics Society 34th Annual Meeting,* p. 623-627.

Lai, J., Wood, D. and Considine, M. (2000). The effect of task conditions on the comprehensibility of synthetic speech. In *CHI Conference Proceedings: Conference on Human Factors in Computing Systems* (pp. 321-328). New York: Association for Computing Machinery.

Lancaster, J.A., Saleem, J.J., Robinson, G.A., Kleiner, B.K., Casali, J.G. (2003). Preliminary study on the effect of approach angle and lower landing minimum level on pilot performance in a low-fidelity static aircraft simulator. Human Factors Engineering and Ergonomics Center, Virginia Tech. In *Proceedings of the 12th International Symposium on Aviation Psychology*, April 14-17, 2003 in Dayton, Ohio USA.

Lassiter, D.L., Morrow, D.G., Hinson, G.E., Miller, M., Hambrick, D.Z. (1996). Expertise and Age Effects on Pilot Mental Workload in a Simulated Aviation Task. In *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting,* pp. 133-137.

Latorella, K. (1998). Effects of modality on interrupted flight deck performance implications for data link. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting,* pp. 87-91. Santa Monica, CA: Human Factors and Ergonomics Society.

Lave, J. (1988). Cognition in practice: mind, mathematics and culture in everyday life. New York: Cambridge University Press.

Lee, A.G. and Simpson, C. (1998). Rapid pilot interface development simulator. In *American Helicopter Society 54th Annual Forum*, pp. 765-776.

Liu, Y. (2001). Comparative study of the effects of auditory, visual and multimodality displays on drivers' performance in advanced traveler information systems. Ergonomics, 44, 4, pp. 425-442.

Liu, Y., Schreiner, C.S., Dingus, T.A. (2000). The effect of advanced traveler information display modality on driver performance. In *Proceedings of the IEA 2000/HFES 2000 Congress*, p.3-234-3-237.

Maggart and Hubal (1998). Situational Awareness Requirements Workshop. In Gawron, V. (Ed.), Guide for Selecting Workload and Situational Awareness Measures Workshop. Human Factors and Ergonomics Society 46th Annual Meeting, Baltimore, MD.

Martin, L. and Flin, R. (1997). Building fire commander situational awareness from team shared mental models. In *Proceedings of the IEEE Colloquium on Computer Mediated*, pp. 2/2-2/7.

Marshall, S.P. (1996). Monitoring situational awareness in tactical information displays with eye-tracking instrumentation. Technical Report (N00014-95-1-1091). San Diego: San Diego State University, August 1996.

Meister, D. (1995). Divergent viewpoints: Essays on human factors questions, 1995.

Melanson, D., Curry, R.E., Howell, J.D. and Connelly, M.E. (1973, May). The effect of communications and traffic situation displays on pilots' awareness of traffic in the terminal area. Presented at the 9th Annual Conference on Manual Control, Massachusetts Institute of Technology.

Metalis, S.A. (1993). Assessment of Pilot Situational Awareness: Measurement via Simulation. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting,* pp. 113-7.

Momtahan, K.L., (1990). Mapping of psychoacoustic parameters to the perceived urgency of auditory warning signals. Unpublished Master's thesis, Carleton University, Ottowa, Ontario.

Muniz, E.J., Stout, R.J., Bowers, C.A., and Salas, E. (1993). A methodology for measuring team situational awareness linked indicators adapted to novel tasks (SALIANT). NASA report no. 19990018345.

Nilsson, M, Soli, S.D., Sullivan, J.A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. Journal of the Acoustical Society of America, 95, 2, pp. 1085-1099.

Nisbett, R.E. and Wilson, T.D. (1977). Telling more than we can know: verbal reports on mental processes. Psychological Review, 84, 3, pp. 231-259.

Paris, C.R., Thomas, M.H., Gilson, R.D. and Kincaid, J.P. (2000). Linguistic cues and memory for synthetic and natural speech. Human Factors, 42, 3, pp. 421-431.

Phillips, E.H. (1992). Langley Data Link System Provides ATC Communications, Weather Depiction. Aviation Week and Space Technology, 136, 1, 52-53.

Prinzo, O.V. (1996). An Analysis of Approach Control/Pilot Voice Communications. *FAA Civil Aeromedical Institute,* (DOT/FAA/AM-96/26), p. 3-44.

Prinzo, O.V. (1999). Voice Communication in a Simulated Approach Control Environment. In *Proceedings of the 10th International Symposium on Aviation Psychology,* p. 603-608.

Regal, D.M., Rogers, W.H., and Boucek, G.P. (1988). Situational awareness in the commercial flight deck: definition, measurement, and enhancement. (SAE Tech. Report SAE 881508). Aerospace Technology Conference and Exposition, Boeing.

Rehmann, A.J. (1993). Airborne data link operational evaluation test plan (Tech. Report DOT/FAA/CT-TN93/30). Atlantic City International Airport, NJ: William J. Hughes Technical Center.

Rehmann, A.J., Reynolds, M.C., and Naumeier, M.E. (1993). FAA airborne data link human factors research plan (Tech. Report DOT/FAA/CT-TN93/5). Atlantic City International Airport, NJ: William J. Hughes Technical Center.

Rehmann, A.J. (1996). Airborne data link study report (Tech. Report DOT/FAA/CT-TN95/62). Pleasantville, NJ: Computer Technology Associates, Inc.

Rehmann, A.J. (1997). Human factors recommendations for airborne controller-pilot data link communication (CPDLC) systems: a synthesis of research results and literature (Tech. Report DOT/FAA/Ct-TN97/6). Atlantic City International Airport, NJ: William J. Hughes Technical Center.

Reynolds, M.E., Fucci, D. and Bond, Z.S. (1997). Effect of visual cuing on synthetic speech intelligibility: a comparison of native and nonnative speakers of English. Perceptual and Motor Skills 84, pp. 695-698.

Ricard, G.L. and Meirs, S.L. (1994). Intelligibility and localization of speech from virtual directions. Human Factors, 36, 1, pp. 120-128.

Roscoe, S. N. (1980). Aviation Psychology, The Iowa State University Press, Ames, Iowa.

Roscoe, S.N., Corl, L. and LaRoche, J. (1997). Predicting human Performance. In Gawron (Ed.), Guide for Selecting Workload and Situational Awareness Measures Workshop. Human Factors and Ergonomics Society 46th Annual Meeting, Baltimore, MD.

Sarter, N.B. and Woods, D.D. (1991). Situation awareness: a critical but ill-defined phenomenon. The International Journal of Aviation Psychology, 1, 1, pp. 45-57.

Salvendy, G. (Ed.). (1997). Handbook of Human Factors and Ergonomics. New York: John Wiley & Sons, Inc.

Sanders, M.S. and McCormick, E.J., (1993). Human Factors in Engineering and Design, McGraw-Hill, Inc., New York, NY.

Selcon, S.J. and Taylor, R.M. (1990). Evaluation of the situational awareness rating technique (SART) as a tool for aircrew systems design. In Situational Awareness in Aerospace Operations (AGARD-CP-478) (pp. 5/1-5/8.

Selcon, S.J., Taylor, R.M. and Koritsas, E. (1991). Workload or situational awareness?: TLX vs. SART for aerospace systems design evaluation. In *Proceedings of the Human Factors and Ergonomics Society 35th Annual Meeting* (pp. 62-66). Santa Monica, CA: Human Factors and Ergonomics Society.

Shannon, C.E., and Weaver, W. (1949). The mathematical theory of communication. Urbana: University of Illinois Press.

Shook, R.W.C., Bandiero, M., Coello, J.P., Garland, D.J., Endsley, M.R. (2000). Situation awareness problems in general aviation. In *Proceedings of the 14th Triennial Congress of the International Ergonomics Association and the 44th Annual Meeting of the Human Factors and Ergonomics Society*, p. 540-544.

Simpson, C.A. and Marhionda-Frost, K. (1984). Synthesized Speech Rate and Pitch Effects on Intelligibility of Warning Messages for Pilots. Human Factors, 26, 5, 509-517.

Simpson, C.A., Marchionda-Frost, K., Navarro, T. (1984). Comparison of Voice Types for Helicopter Voice Warning Systems. In *Proceedings of the 3rd Aerospace Behavioral Engineering Technology Conference*, SAE, Warrendale, PA, 217-225.

Simpson, C.A., McCauley, M.E., Roland, E.F., Ruth, J.C., and Williges, B.H. (1985). System design for speech recognition and generation. Human Factors, 27, pp. 115-141.

Simpson, C.A. and Williams, D.H. (1980). Response time effects of alerting tone and semantic context for synthesized voice cockpit warnings. Human Factors, 22, pp. 319-330.

Smith, K. and Hancock, P.A. (1995). Situation awareness is adaptive, externally directed consciousness. Human Factors, 37, 1, pp. 137-148.

Sorkin, R.D. (1987). Design of auditory and tactile displays. In G. Salvendy (Ed.), Handbook of Human Factors, New York: Wiley and Sons.

Speaks C., Karmen, J.L., and Benitez, L. (1967). Effect of a competing message on synthetic sentence identification. Journal of Speech and Hearing Research, 10, pp. 115-141.

Speech Synthesis. (2002). All-prosodic speech synthesis architecture. Retrieved on June 13, 2002, from: http://www.phon.ox.ac.uk/~jcoleman/IPOX/ipox.html

Stager, P. (2000). Achieving the objectives of certification through validation: methodological issues. In J.A. Wise and V.D. Hopkin (Eds.), Human Factors in Certification (pp. 91-104). New Jersey: Lawrence Erlbaum Associates.

Stern, S.E., Mullenix, J.W., Dyson, C., and Wilson, S.J. (1999). The persuasiveness of synthetic speech versus human speech. Human Factors, 41, 4, pp. 588-595.

Stokes, A.F. and Downs, J. (1998). CRT Text vs. VHF Voice: Effects of Communication Modality in Single-Pilot Cockpits. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, pp. 414-418.

Stokes, A., Wickens, C., and Kite, K. (1990). Display Technology Human Factors Concepts. Society of Automotive Engineers, Inc., Warrendale, PA.

Svensson, E. (1997). Pilot Mental Workload and Situational Awareness—Psychological Models of the Pilot. In Decision Making Under Stress—Emerging Themes and Applications, (pp. 260-7). Ashgate, Brookfield, USA.

Svensson, E., Angelborg-Thanderz, M., Sjoberg, L. (1993). Mission challenge, mental workload and performance in military aviation. Aviation, Space, and Environmental Medicine, November, pp. 985-91.

Talx. (2002). Talx: Keys to High-Quality Digitized Speech. Retrieved June 14, 2002, from: http://www.talx.com/new/technologies/vtvoice.htm

Taylor, R.M. (1990). Situational awareness rating technique (SART): the development of a tool for aircrew systems design. In Situational awareness in aerospace operations (AGARD-CP-478) (pp. 3/1-3/17). Neuilly Sur Seine, France: NATO-AGARD.

Thompson, J. G. and Zhang, Xu (2000). Coordinated Flight Along a Complex Flight Path. In *Proceedings of the 19th Digital Avionics Systems Conference—Entering the Second Century of Powered Flight*, 2 (A01-27744 05-01), p. 5.B. 4-1 to 5.B. 4-6.

Townsend, S. (1992). Report on Airline Fleet Composition (in Preparation), CTA, INCORPORATED, Pleasantville, NJ, 1992.

Trollip, S.R. and Jensen, R.S. (1991). Human Factors for General Aviation. Englewood, CO: Jeppesen Sanderson.

Tsimhoni, O., Green, P. and Lai, J. (2001). Listening to natural and synthesized speech while driving: effects on user performance. International Journal of Speech Technology, 4, 2, pp. 155-169.

US Department of Transportation (2000). Actions to reduce operational errors and deviations have not been effective. *Audit Report AV-2000-01.*

Veltman, J.A., Gailliard, A.W.K. (1993). Measurement of Pilot Workload with Subjective and Physiological Techniques. In *Proceedings of the Workload Assessment and Aviation Conference,* pp. 3.1-3.13.

Venkatagiri, H.S. (2003). Segmental intelligibility of four currently used text-to-speech synthesis methods. Journal of the Acoustical Society of America, 113, 14, p. 2095-2014.

Vidulich, M.A. and Hughes, E.B. (1991). Testing a subjective metric of situation awareness. In *Proceedings of the Human Factors and Ergonomics Society 35th Annual Meeting*, (2) pp. A93-27126 09-54.

Walker, J., Alicandri, E., Sedney, C., and Roberts, K. (1991). In-vehicle navigation devices: effects on the safety of driving performance. Tech Report FHWA-RD-90-053. Department of Transportation.

Wickens, C.D., Sandry, D., and Vidulich, M. (1983). Short-term memory for verbal glance route information. In *Proceedings of the Human Factors Society 23rd Annual Meeting,* Santa Monica, CA: Human Factors Society.

Wickens, C.D. and Hollands, J.G. (2000). Engineering Psychology and Human Performance. New Jersey: Prentice Hall.

Wierwille, W.W. (1979). Physiological measures of aircrew mental workload. Human Factors 21, pp. 575-593.

Wiener, E. L. and Nagel, D. C., (1988). Human Factors in Aviation, Academic Press, Inc., San Diego, CA.

Wierwille, W.W. and Casali, J.G. (1983). A validated rating scale for global mental workload measurement applications. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 129-133). Santa Monica, CA: Human Factors and Ergonomics Society.

Wierwille, W.W. and Eggemeier, F.T (1993). Recommendations for Mental Workload Measurement in a Test and Evaluation Environment. Human Factors, 35, 2, pp. 263-81.

Wierwille, W.W., Rahimi, M., Casali, J.G. (1985). Evaluation of 16 Measures of Mental Workload Using a Simulated Flight Task Emphasizing Mediational Activity. Human Factors, 27, 5, pp. 489-502.

Willits, P. (ed.)(2002). Private Pilot Manual. Jeppesen Sanderson, Inc., Canada.

Wilson, G.F., Badeau, A. (1992). Psychophysiological Measures of Cognitive Workload in Laboratory and Flight. In *Proceedings of the 6th Annual Workshop on Space Operations Applications and Research (SOAR),* p. 474-81.

Wise, J.A., Hopkin, V.D., Garland, D.J., (1994). Human Factors Certification of Advanced Aviation Technologies, Embry-Riddle Aeronautical University Press, Daytona Beach, FL.

Wise, J.A., Hopkin, V.D., Smith, M.L., (1991). Automation and Systems Issues in Air Traffic Control, Springer-Verlag, Berlin: Germany.

APPENDICES

APPENDIX A – Assorted Images of Experimental Apparatus

The Integrated General Aviation Training Environment (i-GATE) Flight Simulator (note the location of the activated visual alerting light just above and to right of screen)

Experimenter's control station.

Sound equipment used for the amplification of aircraft engine noise and for control of MRT stimuli.

Infiniti SM-155 Loudspeakers used to present aircraft engine noise.

Artificial head used to verify headset sound levels and for engine noise calibration. Pictured with Bose® ANR aviation headset.

Flight experiment setup. The touchscreen PC upon which textual ATC messages were presented is pictured at right.

APPENDIX B—Informed Consent Form for the Speech Intelligibility Experiment

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Informed Consent for Participants in Research Projects Involving Human Subjects

Title of Project: Speech Intelligibility of DECtalk vs. AT&T's Natural Voices

Prinicpal Investigator: Jeff A. Lancaster, M.S.

Faculty Advisor: Dr. John G. Casali, Professor, ISE

I. THE PURPOSE OF THIS RESEARCH

The purpose of this study is to determine which of two TTS (text-to-speech) speech synthesis software suites is superior with respect to speech intelligibility. The results of the study will provide researchers with valuable data outlining the intelligibility of speech using these two systems.

II. PROCEDURES

The procedures used in this research are as follows. If you wish to become a participant after reading the description of the study, then sign this form. If you have any questions about the study or this form, please feel free to ask them at any time.

The study consists of three sessions. For the first session, you will be screened to determine if you qualify for the experiment. Screening will consist of a hearing test and several assessment tests. To begin with, you will be asked several questions to assess the general health and condition of your ears. Then you will be given an examination in which the experimenter will look into your ears using an otoscope. Next, your right and left hearing will be tested with very quiet tones played through a set of headphones. You will have to be very attentive and listen carefully for these tones. Depress the button on the hand-held switch and hold it down whenever you hear the pulsed-tones and release it when you do not hear the tones. The tones will be very faint and you will have to listen carefully to hear them. No loud or harmful sounds will be presented over the headphones.

If you qualify and choose to participate in the study, you will be instructed in and allowed to practice the procedures used in the modified rhyme test (MRT). First, you will be given a list of 300 words, which are divided into 50 groups of six words each and asked to read and familiarize yourself with the words on the list. (You may ask the experimenter to show you an example of this list now if you would like to see it.) The experimenter will be in another room and will control the presentation of the speech and noise. A total of 50 target words, one word from each six-word group, will be presented within the carrier sentence: "Mark the word _____ now." The pre-recorded sentences will be presented through a Bose® Active-noise Reduction aviation headset. During the test, you should circle or otherwise mark each target word spoken. The sentences/words

may be difficult to understand, so you must concentrate on the task and listen intently for the sentences/words. After all 50 sentences have been presented, the experimenter will examine your data and determine if additional practice trials are necessary. At a minimum, this practice session will consist of a complete list of 50 words, but additional practice trials may be conducted if you or the experimenter thinks they are necessary.

If you choose to continue with the experiment, you will be asked to participate in two experimental sessions. In these experimental sessions, additional modified rhyme tests will be conducted in a manner identical to that described above. The test sessions will take place on separate days, with each session evaluating one of the synthesizers. The standard governing this procedure (ANSI 3.2-1989) requires at least two trials at each S/N level to produce an average, so each session will require six trials (two for each of the three S/N ratios). Participants will be asked to leave the experimental room whenever the experimenter sets, adjusts, and verifies the speech level under the headset. Each experimental session will take approximately one hour.

III. RISKS

During the hearing test, you will be in a soundproof booth with the experimenter sitting outside. The door to the booth will be shut but not locked; you may open it from the inside or the experimenter may open it from the outside. There is also an intercom system through which you may communicate with the experimenter by simply talking (there are no buttons to push). If you are or think you may be claustrophobic or if you are uncomfortable in the confined spaces, please tell the experimenter at this time. He/she will show you the rooms and let you enter them to see if they make you uncomfortable. The speech intelligibility test will be conducted in another room and the experimenter will be in an adjacent room.

The Occupational Safety and Health Administration (OSHA) currently allows workers in the United States to be exposed to 90 dBA time-weighted average noise for 8 hours/day. Sound measurements have been conducted using an artificial head (ANSI 3.19-1974) within the aircraft noise (85 dBA) to be utilized in the experiment to determine the sound pressure levels (SPL) under the Bose® Active-noise Reduction headset. The results indicate a SPL of 64 dBA under the headset. Speech-to-noise ratios (S/N, the difference between the speech level and noise level) will affect your performance on the modified rhyme test. Different S/N ratios will be utilized to see how intelligibility is affected. At no time will the S/N ratio exceed 15 dB. Even at the highest S/N ratio of 15 dB, at no time will the stimuli presented to you under the headset exceed 80 dBA, which is well below the Occupational Safety and Health Administration's (OSHA) 8-hour exposure limit for an employee's workday (i.e., 90 dBA).

Given the short exposure times, it is felt that there is little or no potential for doing any harm to your hearing. (Stimulus levels presented during the experiment will be checked and adjusted before every experimental session.)

III. BENEFITS OF THIS PROJECT

Your participation in this experiment will provide information that will be used to determine the relative strengths and weaknesses of two current TTS speech synthesis software titles. The results of this study will help to determine which TTS engine is superior for use in human/machine systems.

IV. EXTENT OF ANONYMITY AND CONFIDENTIALITY

The results of this study will be kept strictly confidential. At no time will the researchers release the results of the study to anyone other than the individuals working on the project without your written consent. Your written consent is required for the researcher to release any data identified with you as an individual to anyone other than personnel working on the project. The information you provide will have your name removed and only a subject number will identify you during analyses and any written reports of the research. All subject numbers will be secure and stored on the principal investigator's personal computer.

VI. COMPENSATION

You will be paid $8.00 per hour for your participation in the experiment. Payment will be made immediately after you have finished your participation.

VII. FREEDOM TO WITHDRAW

You are free to withdraw from this study at any time without penalty. If you choose to withdraw, you will be compensated for the portion of time you have spent in the study. There may also be certain circumstances under which the investigator may determine that you should not continue as a participant of this project. These include, but are not limited to, unforeseen health-related difficulties, inability to perform the task, and unforeseen danger to the participant, experimenter, or equipment.

VIII. APPROVAL OF RESEARCH

This research has been approved, as required, by the Institutional Review Board (IRB) for projects involving human subjects at the Virginia Polytechnic Institute and State University and by the Grado Department of Industrial and Systems Engineering.

_____ _____

IRB Approval Date Approval Expiration Date

IX. PARTICIPANT'S RESPONSIBILITIES AND PERMISSION

I voluntarily agree to participate in this study and I know of no reason why I cannot participate. I have read and understand the informed consent and conditions of this project, and understand that I have the following responsibilities: (1) to listen attentively to the stimulus sounds presented during the tests, to respond appropriately and accurately, and to follow all instructions to the best of my ability, (2) to notify the experimenter at any time about discomfort or a desire to discontinue participation. I hereby acknowledge the above and give my voluntary consent for participation in this project. If I participate, I may withdraw at any time without penalty.

_____ _____

Print Name Subject's Signature Date

Should I have any pertinent questions about this research or its conduct, and research subjects' rights, and whom to contact in the event of a research-related injury to the subject, I may contact:

Jeff A. Lancaster (Principal Investigator) - 540-231-9086 / jlancast@vt.edu

Dr. John G. Casali, (Faculty Advisor) - 540-231-9081 / jcasali@vt.edu

Dr. Gary S. Robinson, (Faculty Advisor) - 540-231-2680 / grobins@vt.edu

Dr. Brian M. Kleiner, (Faculty Advisor) - 540-231-4926 / bkleiner@vt.edu

David M. Moore 540-231-4991 / moored@vt.edu
Chair, IRB
Office of Research Compliance
Research & Graduate Studies

This Informed Consent is valid from __June 4, 2003__ to __July 4, 2004__.

[NOTE: Subjects will be given a complete copy (or duplicate original) of the signed Informed Consent.]

SCREENING FORM

Pure-Tone Audiometric Tests for Normal Hearing

Participant: _____ Age:_____ Sex:_____

Phone:_____ Screening Date:_____ Qualify?____

Right Ear

Frequency (Hz)	t-1	t-2	t-3	t-4	t-5	t-6	final threshold
125	_____	_____	_____	_____	_____	_____	_____
250	_____	_____	_____	_____	_____	_____	_____
500	_____	_____	_____	_____	_____	_____	_____
1000	_____	_____	_____	_____	_____	_____	_____
2000	_____	_____	_____	_____	_____	_____	_____
3000	_____	_____	_____	_____	_____	_____	_____
4000	_____	_____	_____	_____	_____	_____	_____
6000	_____	_____	_____	_____	_____	_____	_____
8000	_____	_____	_____	_____	_____	_____	_____

Left Ear

Frequency (Hz)	t-1	t-2	t-3	t-4	t-5	t-6	final threshold
125	_____	_____	_____	_____	_____	_____	_____
250	_____	_____	_____	_____	_____	_____	_____
500	_____	_____	_____	_____	_____	_____	_____
1000	_____	_____	_____	_____	_____	_____	_____
2000	_____	_____	_____	_____	_____	_____	_____
3000	_____	_____	_____	_____	_____	_____	_____
4000	_____	_____	_____	_____	_____	_____	_____
6000	_____	_____	_____	_____	_____	_____	_____
8000	_____	_____	_____	_____	_____	_____	_____

SCREENING FORM

Otoscopic Data

Occluding wax?: _____

Ear canal irritation?: _____

Unusual canal characteristics: _____

Eardrum perforations?: _____

Eardrum scar tissue? _____

Foreign matter?: _____

Self-Report Data

Tinnitus or head noises: _____

Otopathological history: _____

Occupation: _____

Noisy hobbies: _____

HPD experience: _____

Other: _____

APPENDIX C—Modified Rhyme Test (MRT) Word List

1	went	sent	bent	14	not	tot	got	27	peel	reel	feel
	dent	tent	rent		pot	hot	lot		eel	keel	heel
2	hold	cold	told	15	vest	test	rest	28	hark	dark	mark
	fold	sold	gold		best	west	nest		bark	park	lark
3	pat	pad	pan	16	pig	pill	pin	29	heave	hear	heat
	path	pack	pass		pip	pit	pick		heal	heap	heath
4	lane	lay	late	17	back	bath	bad	30	cup	cut	cud
	lake	lace	lame		bass	bat	ban		cuff	cuss	cub
5	kit	bit	fit	18	way	may	say	31	thaw	law	raw
	hit	wit	sit		pay	day	gay		paw	jaw	saw
6	must	bust	gust	19	pig	big	dig	32	pen	hen	men
	rust	dust	just		wig	rig	fig		then	den	ten
7	teak	team	teal	20	pale	pace	page	33	puff	puck	pub
	teach	tear	tease		pane	pay	pave		pus	pup	pun
8	din	dill	dim	21	cane	case	cape	34	bean	beach	beat
	dig	dip	did		cake	came	cave		beak	bead	beam
9	bed	led	fed	22	shop	mop	cop	35	heat	neat	feat
	red	wed	shed		top	hop	pop		seat	meat	beat
10	pin	sin	tin	23	coil	oil	soil	36	dip	sip	hip
	fin	din	win		toil	boil	foil		tip	lip	rip
11	dug	dung	duck	24	tan	tang	tap	37	kill	kin	kit
	dud	dub	dun		tack	tam	tab		kick	king	kid
12	sum	sun	sung	25	fit	fib	fizz	38	hang	sang	bang
	sup	sub	sud		fill	fig	fin		rang	fang	gang
13	seep	seen	seethe	26	same	name	game	39	took	cook	look
	seek	seem	seed		tame	came	fame		hook	shook	book

40	mass	math	map	41	ray	raze	rate	42	save	same	sale
	mat	man	mad		rave	rake	race		sane	sake	safe
43	fill	kill	will	44	sill	sick	sip	45	bale	gale	sale
	hill	till	bill		sing	sit	sin		tale	pale	male
46	wick	sick	kick	47	peace	peas	peak	48	bun	bus	but
	lick	pick	tick		peach	peat	peal		bug	buck	buff
49	sag	sat	sass	50	fun	sun	bun				
	sack	sad	sap		gun	run	nun				

APPENDIX D—Informed Consent Form for the Data Link Performance Experiment

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Informed Consent for Participants in Research Projects Involving Human Subjects

Title of Project: Investigating Pilot Performance Using Mixed-modality Simulated Data Link

Investigator: Jeff A. Lancaster, M.S.

I. THE PURPOSE OF THIS RESEARCH

The purpose of this study is to assess how flying in a controlled airspace is affected by different presentation formats of ATC directives (e.g., synthesized voice) using a simulated data link. In the study, you will be asked to fly a simulator within a controlled airspace in daytime conditions. The simulator consists of a console with an avionics panel, pilot controls, and a small out-the-window view. The controls have the look and feel of those in a real aircraft, and the aircraft response is derived from accurate flight aerodynamics. The simulator is a Federal Aviation Administration (FAA) certified trainer. Including you, 16 current VFR-rated pilots of at least 18 years of age will be participating in the experiment.

II. PROCEDURES

The study consists of three sessions. Aside from the first half of the first session, wherein your hearing will be tested, you will report to the MacroErgonomics and Group Decision Systems Laboratory (MGDSL), room 567 Whittemore Hall, on the Virginia Tech campus for the study. The first session involves testing your hearing (resulting in an audiogram) and of familiarization with the flight simulator. The audiogram is to ensure that your hearing is adequate such that you can hear auditory alerts. Following successful completion of the audiogram, you will then be introduced to the i-GATE flight simulator, which will enable you to get used to the control and feel of the simulator (familiarization). The first session should take from an hour to an hour and a half to complete.

The second and third sessions involve actual flight-testing, with each session consisting of two separate flight scenarios. You will fly each scenario within a fictitious controlled airspace, starting during cruise and ending with a landing approach. Your task is to respond to messages from ATC. These messages may include vectoring commands, advisories, or other flight information. Your aircraft will be tuned to the frequency of a simulated control tower. You should fly the aircraft and respond to any ATC commands as you normally would during routine operations, except this time you will use a simulated data link. Use of the simulated data link requires you to suspend what may be ingrained or automatic operations on your part, most notably the ‘call back’ or ‘read back’ of the ATC message. Since data link does not rely on traditional ‘party line’

communications (i.e., radio traffic and chatter), there is no controller to respond to—it is a computer. For this study, you must also suspend your operation of the radio stack by assuming all hand-offs and frequency changes (e.g., from tower frequency to ground frequency) occur automatically and do not require manual changes of frequencies on the radio stack. Incoming data link messages from ATC will be alerted to you through a visual alert and auditory alert (a flashing light and bell sound, respectively). When you notice the alert(s), you must press 'RECEIVE MESSAGE' using the simulated data link touchscreen located to the right of the simulator in order to receive the incoming ATC message. The message will then be presented to you through different voice and/or textual formats using an aviation headset and/or on the data link screen itself, depending on the scenario. Once you have heard and/or have read and understood the message, you must reply to ATC using one of two other on-screen buttons: 'WILCO/ROGER' or 'UNABLE' (for 'will comply/roger' or 'unable to comply'), even if the ATC message is an advisory requiring no control input on your part. You may also choose to repeat the message if desired. Once you choose 'WILCO/ROGER' or 'UNABLE' you should then maneuver the aircraft as directed by the ATC message (if applicable). During the flight, the simulation may or may not be interrupted several times and questions will be asked of you related to current flight operations. Upon completion of the flight scenario, you will be asked to rate the difficulty of the flight as well as complete a questionnaire.

III. RISKS

The risks of harm anticipated in the proposed research are not greater, considering the probability and magnitude, than those encountered in daily life.

Sound levels within the laboratory will mirror those of the aircraft being simulated. Subjects will wear circumaural aviation headsets to a) help protect them from noise exposure, and b) to more closely replicate actual aircraft conditions.

III. BENEFITS OF THIS PROJECT

Your participation in this experiment will provide information that may be helpful in understanding pilot performance, workload, and situation awareness in various levels of ceiling/visibility and types of ATC presentation directives. The results of this research may help engineers and researchers in the development of envisioned future general aviation systems by providing data that outlines the effects of specific conditions on pilot performance, workload, and situation awareness. No guarantee of direct benefits has been made to encourage you to participate. If you would like to receive a summary of this research when it is completed, please provide a self-addressed envelope.

IV. EXTENT OF ANONYMITY AND CONFIDENTIALITY

The results of this study will be kept strictly confidential. Your written consent is required for the researcher to release any data identified with you as an individual to anyone other than personnel working on the project. The information you provide will have your name removed and only a subject number will identify you during analyses and any written reports of the research. All subject numbers will be secure and stored on the principal investigator's personal computer.

VI. COMPENSATION

You will be paid $20.00 an hour for your participation in the experiment. Payment will be made immediately after you have finished your participation.

VII. FREEDOM TO WITHDRAW

You are free to withdraw from this study at any time for any reason with no penalty. You will be compensated for your participation up to the point of withdrawal.

VIII. APPROVAL OF RESEARCH

This research has been approved, as required, by the Institutional Review Board (IRB) for projects involving human subjects at the Virginia Polytechnic Institute and State University and by the Department of Industrial and Systems Engineering.

___June 4, 2003___	___July 4, 2004_____
IRB Approval Date	Approval Expiration Date

IX. SUBJECT'S RESPONSIBILITIES AND PERMISSION

I voluntarily agree to participate in this study and I know of no reason why I cannot participate. I have read and understand the informed consent and conditions of this project. I have had all my questions answered. I hereby acknowledge the above and give my voluntary consent for participation in this project. If I participate, I may withdraw at any time without penalty.

_____	_____	
Print Name	Subject's Signature	Date

Should I have any pertinent questions about this research or its conduct, and research subjects' rights, and whom to contact in the event of a research-related injury to the subject, I may contact:

Jeff A. Lancaster (Investigator) - 540-231-9086 / jlancast@vt.edu

Dr. John G. Casali, (Faculty Advisor) - 540-231-9081 / jcasali@vt.edu

Dr. Gary S. Robinson, (Faculty Advisor) - 540-231-2680 / grobins@vt.edu

Dr. Brian M. Kleiner, (Faculty Advisor) - 540-231-4926 / bkleiner@vt.edu

Dr. Thurmon E. Lockhart (Faculty Advisor) - 540-231-9088 / lockhart@vt.edu

Dr. Antonio A. Trani (Faculty Advisory) - 540-231-4418 / vuela@vt.edu

David M. Moore, Chair, IRB 540-231-4991 / moored@vt.edu
Office of Research Compliance
Research & Graduate Studies

This Informed Consent is valid from __June 4, 2003__ to __July 4, 2004__.

[NOTE: Subjects will be given a complete copy (or duplicate original) of the signed Informed Consent.]

APPENDIX E—Subject Data, Data Link Touchscreen, Flight Routes, Data Link Messages

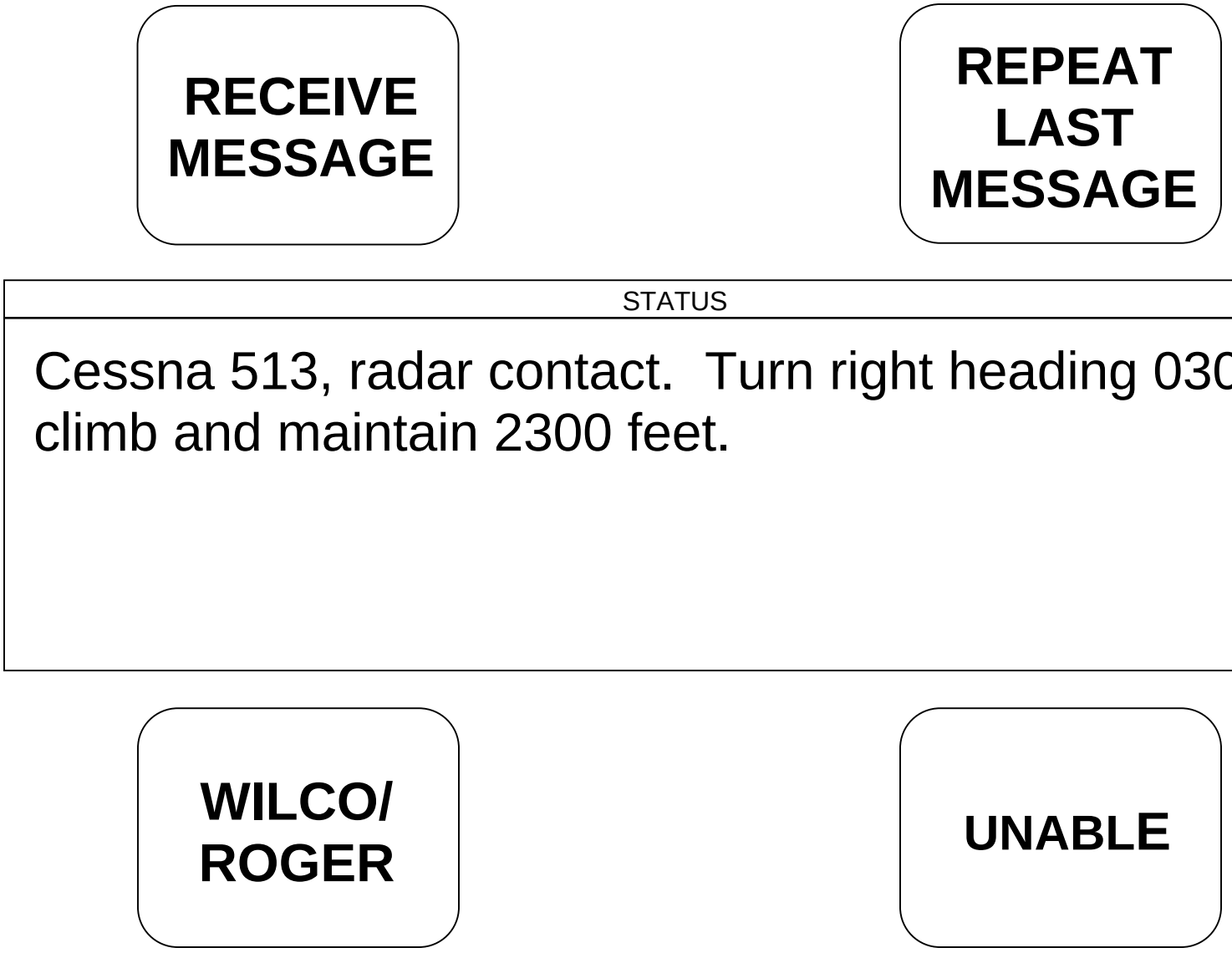
Pilot Audiograms

Subject #	Age	Gender	Ear	125Hz	250Hz	500Hz	1KHz	2KHz	3KHZ	4KHz	6KHz	8KHz
1	21	M	R	10	15	10	5	0	10	20	30	30
			L	10	10	5	0	0	15	15	25	25
2	21	M	R	10	10	5	10	5	0	-10	5	0
			L	5	10	5	5	10	5	-5	0	5
3	70	M	R	15	15	10	0	15	25	35	30	>40
			L	10	10	10	10	15	25	35	>40	>40
4	54	M	R	15	5	5	0	0	5	10	15	-5
			L	10	10	5	0	0	0	20	20	10
5	50	M	R	5	5	10	10	5	15	15	25	25
			L	5	5	10	0	0	115	15	25	20
6	57	M	R	5	5	5	15	20	35	50	>50	>50
			L	0	5	10	15	20	35	40	>50	>50
7	24	M	R	10	5	5	5	-5	-5	-5	0	5
			L	5	5	5	5	-5	-5	10	10	15
8	28	M	R	5	0	0	0	0	15	5	10	15
			L	0	0	0	0	-5	20	5	10	5
9	25	M	R	10	15	5	5	0	5	5	5	5
			L	10	10	10	20	5	0	5	10	5
10	38	M	R	5	-5	-5	-5	-5	10	20	5	15
			L	0	-5	-5	-5	0	5	5	15	15
11	45	M	R	10	15	15	15	15	10	20	25	5
			L	10	15	15	10	15	15	15	10	10
12	47	M	R	10	10	5	5	5	>40	>40	>40	>40
			L	5	5	0	5	5	>40	>40	>40	>40
13	32	M	R	5	10	5	5	10	15	10	20	15
			L	5	10	10	10	5	5	10	20	20
14	22	M	R	5	0	-5	0	5	0	10	15	35
			L	5	10	0	5	5	5	10	10	25
15	47	M	R	15	5	0	-5	0	5	5	0	5
			L	10	5	0	0	10	20	35	10	0
16	35	M	R	5	5	10	10	0	10	5	5	-5
			L	5	5	10	15	5	20	15	5	-5

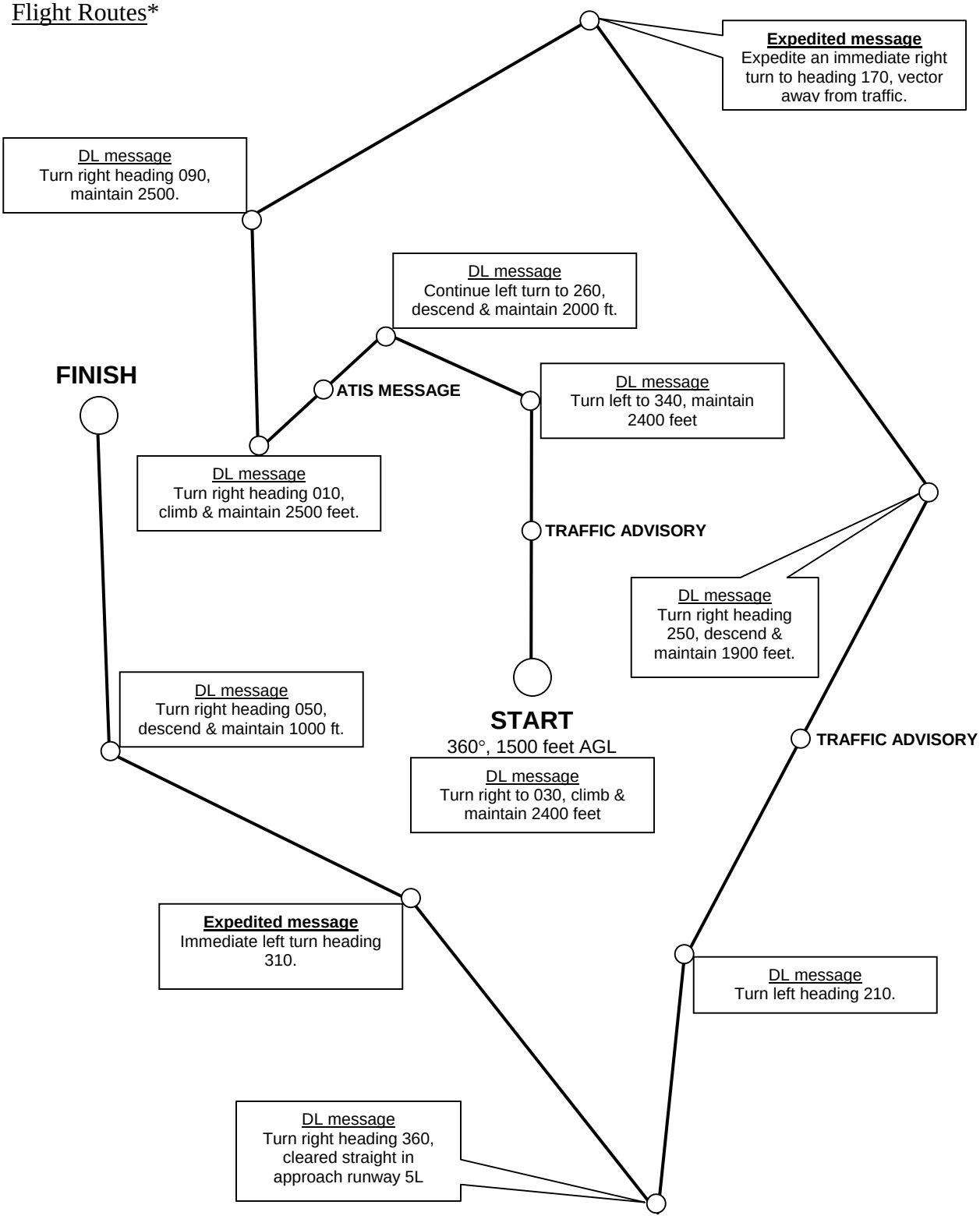
Pilot Audiogram Equalization to MRT Study Participants

Subject #	Most Comfort. Level	TOTAL FLIGHT HOURS	avg. 500-2k	min 500-2k	mean HL MRT - mean HL Pilots	To produce 0 dB S/N	MCL
1	69.3	284	5	1.666666667	-2.833336667	66.83333667	69.3
			1.666666667				
2	65.1	97	6.666666667	6.666666667	-6.666666667	70.66666667	65.1
			6.666666667				
3	69.6	5500	8.333333333	8.333333333	-8.333333333	72.33333333	69.6
			11.66666667				
4	70	3000	1.666666667	1.666666667	-1.666666667	65.66666667	70
			1.666666667				
5	71	800	8.333333333	3.333333333	-3.333333333	67.33333333	71
			3.333333333				
6	67	1200	13.33333333	13.33333333	-13.33333333	77.33333333	67
			15				
7	69.5	100	1.666666667	1.666666667	-1.666666667	65.66666667	69.5
			1.666666667				
8	72	300	0	-1.66666667	1.666666667	65.66666667	72
			-1.666666667				
9	69.9	230	3.333333333	3.333333333	-3.333333333	67.33333333	69.9
			11.66666667				
10	71	280	-5	-5	5	69	71
			-3.333333333				
11	68.9	230	15	13.33333333	-13.33333333	77.33333333	68.9
			13.33333333				
12	66.2	150	5	3.333333333	-3.333333333	67.33333333	66.2
			3.333333333				
13	71	140	6.666666667	6.666666667	-6.666666667	70.66666667	71
			8.333333333				
14	66	250	0	0	0	64	66
			3.333333333				
15	71	480	-1.666666667	-1.66666667	1.666666667	65.66666667	71
			3.333333333				
16	68.7	325	6.666666667	6.666666667	-6.666666667	70.66666667	68.7
			10				

Sample Data Link Touch Screen



Flight Routes*



*Synthesized/text and digitized data links shown. Digitized data link started at finish and ended at start. Image flip, textual data link start to finish; image flip, synthesized data link finish to start. Altitude directives were modified to maintain MVFR conditions as needed.

Aural data link messages

SYNTHESIZED/TEXUAL (START 0-degrees, 1500 ft., 90 kts.)

1. Cessna five one niner, radar contact. Turn right heading zero three zero, climb and maintain two thousand four hundred feet.

2. Cessna five one niner, traffic 11 o'clock, four miles east bound, two thousand six hundred Sundowner.

3. Cessna five one niner, turn left heading three four zero, maintain two thousand four hundred feet. (2300 feet MVFR)

4. Cessna five one niner, continue left turn to two six zero, descend and maintain two thousand feet.

5. Centennial Airport, information Uniform. One one four five Zulu observation. Wind three three zero at eleven. Visibility ten, clear. (Three, 2800 feet MVFR) Temperature two one, check density altitude. Dew point one zero, altimeter three zero point two two. Visual approach in use, landing and departing runways five right and left. Departing runway ten.

6. Cessna five one niner, turn right heading zero one zero, climb and maintain two thousand five hundred feet. (2300 feet MVFR)

7. Cessna five one niner, continue right turn to heading zero nine zero, maintain two thousand five hundred feet. (2300 feet MVFR)

8. Cessna five one niner, ***expedite an immediate right turn to heading one seven zero***, vector away from traffic.

9. Cessna five one niner, turn right heading two five zero, descend and maintain one thousand nine hundred feet.

10. Cessna five one niner, traffic one o'clock three miles east bound, two thousand eight hundred Comanche.

11. Cessna five one niner, turn left heading two one zero.

12. Cessna five one niner, Centennial Tower. Turn right heading three six zero. Cleared straight in approach, runway five left, maintain VFR. (MVFR)

13. Cessna five one niner, ***immediate left turn heading three one zero*.**

14. Cessna five one niner, turn right to heading zero five zero, descend and maintain one thousand feet.

SYNTHESIZED (Start 180-degrees, 1500 ft., 90 kts.)

1. Cessna five one niner, radar contact. Turn right heading two three zero, climb and maintain two thousand three
hundred feet.

2. Cessna five one niner, traffic one o'clock, 5 miles west bound, two thousand seven hundred Lance Air.

3. Cessna five one niner, turn left heading one three zero, maintain two thousand three hundred feet.

4. Cessna five one niner, turn right heading one eight zero, increase airspeed to one hundred knots.

5. Skycoast Airport, information Golf. Ten thirty Zulu observation. Wind one five zero at eleven. Visibility ten, clear. (Three, 2800 feet MVFR). Temperature two two, check density altitude. Dew point one one, altimeter three zero point two one. Visual approach in use, landing and departing runways two one right and left. Departing runway one niner.

6. Cessna five one niner, turn left heading zero three zero, climb and maintain two thousand five hundred feet. (Descend and maintain 2000 feet MVFR).

7. Cessna five one niner, turn right heading zero seven zero, maintain two thousand five hundred feet. (Climb and maintain 2300 feet MVFR).

8. Cessna five one niner, ***immediate left turn to heading three five zero***, vector away from traffic.

9. Cessna five one niner, turn left heading two seven zero, descend and maintain two thousand two hundred feet.

10. Cessna five one niner, traffic ten o'clock, 3 miles eastbound, two thousand eight hundred Tampico.

11. Cessna five one niner, turn left heading one nine zero, descend and maintain two thousand feet. Decrease airspeed to ninety knots.

12. Cessna five one niner, ***expedite left turn heading zero eight zero*.**

13. Cessna five one nine, SkyCoast tower. Turn right heading one six zero, descend and maintain one thousand five hundred feet. Cleared straight in approach, runway two one right, maintain VFR.

14. Cessna five one nine, turn right heading two one zero, descend and maintain one thousand feet.

DIGITIZED (START 0-degrees, 1500 ft., 90 kts.)

1. Cessna five one niner, radar contact. Turn left heading three three zero, climb and maintain two thousand four hundred feet. (2300 feet MVFR)

2. Cessna five one niner, traffic two o'clock four miles westbound, one thousand nine hundred Cherokee.

3. Cessna five one niner, turn right heading zero two zero, increase airspeed to one hundred knots.

4. Cessna five one niner, turn right heading one zero zero, maintain two thousand four hundred feet. (2300 feet MVFR).

5. Bayside Airport, information Delta. Twelve forty-five Zulu observation. Wind one two zero at fifteen. Visibility ten, clear. (Three, 2800 feet MVFR). Temperature two one, check density altitude. Dew point one one, altimeter three zero point one three. Visual approach in use, landing and departing runways three one right and left. Departing runway two six.

6. Cessna five one niner, turn left heading three five zero, increase airspeed to one hundred and ten knots.

7. Cessna five one niner, continue left turn to heading to two seven zero, climb and maintain two thousand five hundred feet. (Descend, maintain 2000 feet MVFR)

8. Cessna five one niner, ***immediate left turn to heading one nine zero***, descend and maintain two thousand two hundred feet. (Maintain 2000 feet MVFR).

9. Cessna five one niner, turn left to heading one one zero, maintain two thousand two hundred feet. (Climb and maintain 2200 feet MVFR).

10. Cessna five one niner, traffic one o'clock, 4 miles westbound, two thousand seven hundred Beechcraft.

11. Cessna five one niner, turn left heading three six zero, descend and maintain one thousand five hundred feet. Decrease airspeed to one hundred knots.

12. Cessna five one niner, Bayside tower. Turn left heading three zero zero, maintain one thousand five hundred feet, decrease airspeed to ninety knots. Vectored for runway three one, final approach course, maintain VFR.

13. Cessna five one niner, ***expedite right turn heading zero now*.**

14. Cessna five one niner, turn left heading three one zero, descend and maintain one thousand feet.

APPENDIX F—Questionnaire, SA Queries, and Rating Scales

Data Link Questionnaire

Subject # _____ Data link condition: _____ Date: _____

1. Did you experience any problems regarding data link operation? If so, please explain:

2. Did you misunderstand anything regarding the test organization and task? If so, please explain:

3. If applicable, please describe the system's voice in your own words, including your impressions of the voice when compared to natural speech.

4. If applicable, please rate the following characteristics of the speech you heard in this scenario by placing an 'X' above the number that most closely corresponds to your impressions:

   The pronunciation was clear:

   1 (strongly disagree) 2 (disagree) 3 (neutral) 4 (agree) 5 (strongly agree)

   Stress was applied appropriately on words that required it:

   1 (strongly disagree) 2 (disagree) 3 (neutral) 4 (agree) 5 (strongly agree)

   The speed of utterances was satisfactory:

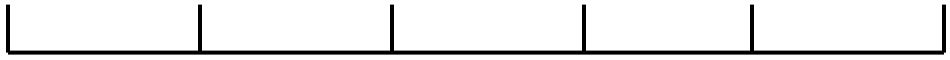
   

   1 (strongly disagree) 2 (disagree) 3 (neutral) 4 (agree) 5 (strongly agree)

The intelligibility of utterances was satisfactory:

1 (strongly disagree) 2 (disagree) 3 (neutral) 4 (agree) 5 (strongly agree)

The utterances were natural-sounding:

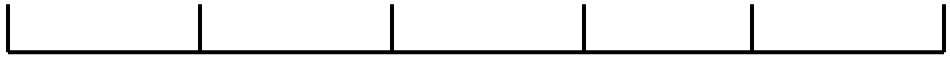


1 (strongly disagree) 2 (disagree) 3 (neutral) 4 (agree) 5 (strongly agree)

Articulation of words was satisfactory:

1 (strongly disagree) 2 (disagree) 3 (neutral) 4 (agree) 5 (strongly agree)

The loudness of utterances was satisfactory:



1 (strongly disagree) 2 (disagree) 3 (neutral) 4 (agree) 5 (strongly agree)

5. Finally, please provide any other comments that you wish to share:

SAGAT Query List

Subject # _____

Query #	SA Level	SAGAT Query	Limits
1	2	What is your current deviation from your intended/assigned heading?	(+/- 10 knots)
2	1	What is your altitude?	(+/- 100 feet)
3	1	What is your heading?	(+/- 10 degrees)
4	3	What is the trajectory of the last traffic advisory relative to ownship?	(Correct or incorrect)
5	1	What was the position of the aircraft during the last traffic advisory?	(Correct or incorrect)
6	2	What is your airspeed?	(Correct or incorrect)
7	1	What was the altimeter setting of the last ATIS message?	(Correct or incorrect)
8	1	What is the active landing runway?	(Correct or incorrect)
9	1	What is the active departing runway?	(+/- 10 degrees)
10	2	What is your current deviation from your intended/assigned altitude?	(+/- 100 feet)
11	3	Assuming no change in heading, in which section of the moving map will the aircraft in the last traffic advisory be located?	(Correct or incorrect)
12	1	What kind of aircraft was the last traffic advisory?	(Correct or incorrect)

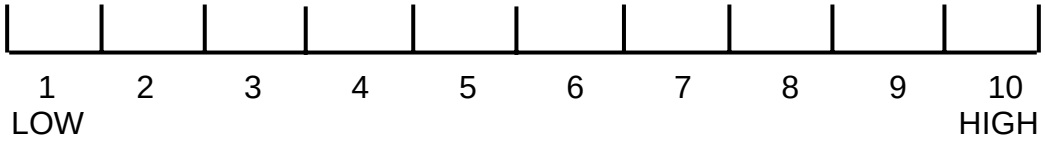
Query #	Textual	SynText	Digitized	Synthesized
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12 DATE RUN				

SART Scale

SART **S#** **DL**

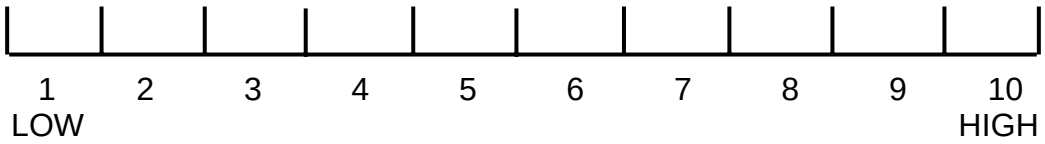
Instability of Situation

How changeable is the situation? Is the situation highly unstable and likely to change suddenly (high), or is it very stable and straightforward (low)?



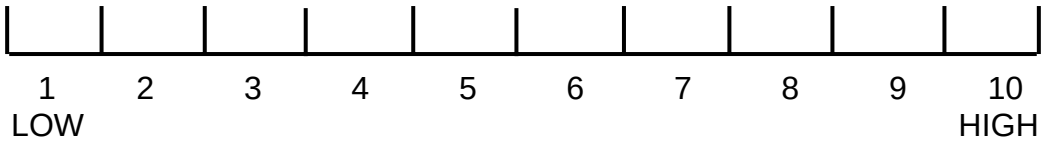
Complexity of Situation

How complicated is the situation? Is it complex with many interrelated components (high) or is it simple and straightforward (low)?



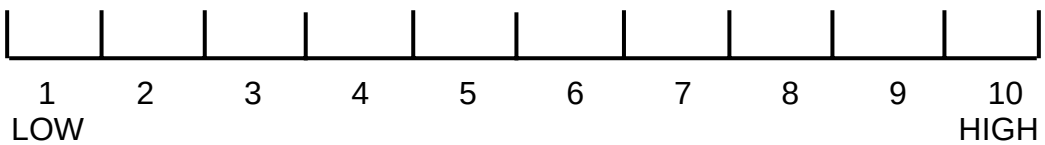
Variability of Situation

How many variables are changing in the situation? Are there a large number of factors varying (high) or are there very few variables changing (low)?



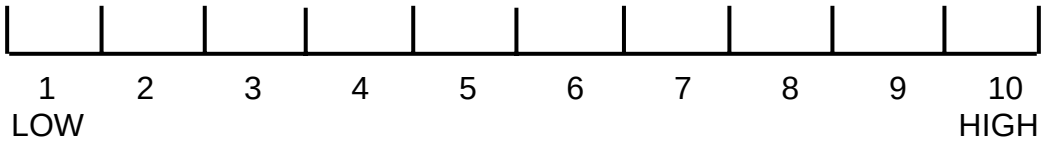
Arousal

How aroused are you in the situation? Are you alert and ready for activity (high) or do you have a low degree of alertness (low)?



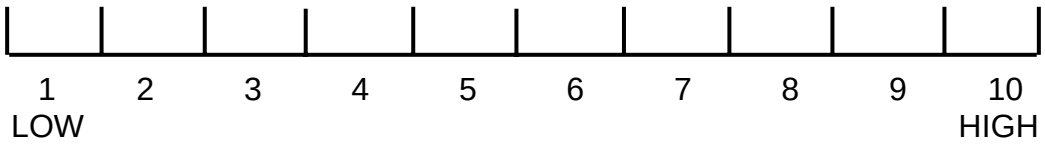
Concentration of Attention

How much are you concentrating on the situation? Are you bringing all your thoughts to bear (high) or is your attention elsewhere (low)?



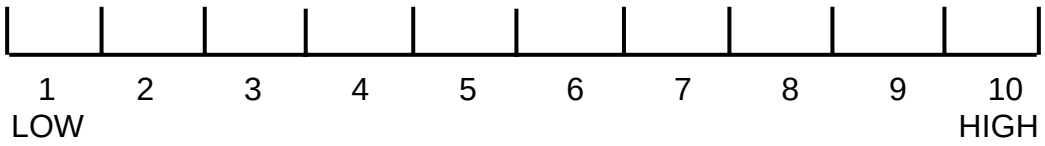
Division of Attention

How much is your attention divided in the situation? Are you concentrating on many aspects of the situation (high) or focused on only one (low)?



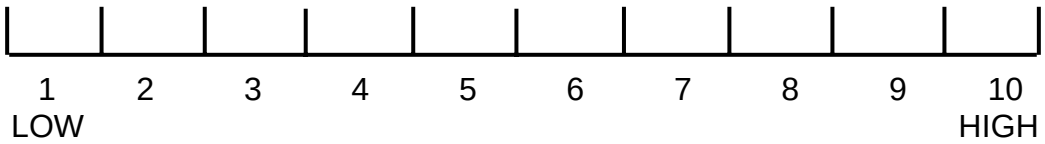
Spare Mental Capacity

How much mental capacity do you have to spare in the situation? Do you have sufficient to attend to many variables (high) or nothing to spare at all (low)?



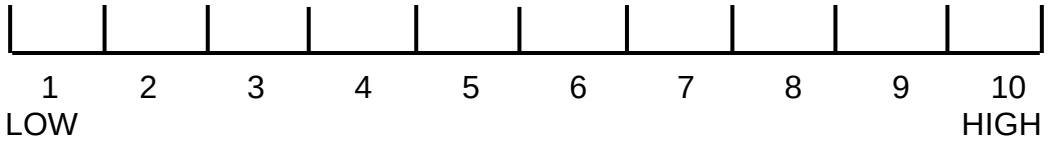
Information Quantity

How much information have you gained about the situation? Have you received and understood a great deal of knowledge (high) or very little (low)?



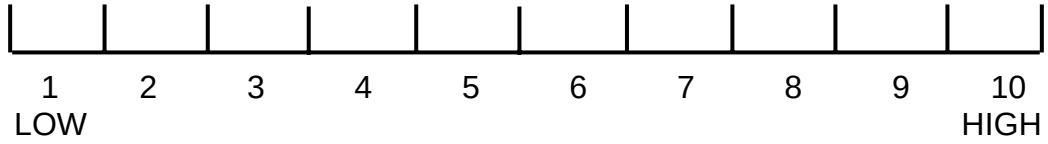
Information Quality

How good is the information you have gained about the situation? Is the knowledge communicated very useful (high) or is it a new situation (low)?



Familiarity with Situation

How familiar are you with the situation? Do you have a great deal of relevant experience (high) or is it a new situation (low)?



APPENDIX G—List of Acronyms

List of Acronyms

AAS	Advanced Automation System
ACT	Auditory Canal Temperature
ADS	Automatic Dependent Surveillance
ADS-B	Automatic Dependent Surveillance-Broadcast
ADO	Assistant Divisional Office
AERA	Automated Enroute ATC Automation System
ANOVA	Analysis of Variance
ANR	Active Noise Reduction
ANSI	American National Standards Institute
API	Application Program Interface
ATA	Air Transport Association
ATC	Air Traffic Control
ATCS	Air Traffic Control Specialist
ATIS	Automated Terminal Information Service
ATM	Air Traffic Management
CPDLS	Controller-Pilot Data Link System
CCC	Command & Control Center
CDU	Control Display Unit
CLSA	China Lake Situation Awareness
CMT	Continuous Memory Test
CNS	Central Nervous System
CRT	Cathode-Ray Tube
DAT	Digital Audio Tape
dB	Decibel
DCT	Duty Cycle Time
DRT	Diagnostic Rhyme Test
DTMF	Dual-Tone Multiple Frequency
EEG	Electroencephalogram
ELS	Electronic Library System
EOG	Electrooculographic
FAA	Federal Aviation Administration
FMS	Flight Management System
GA	General Aviation
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GS	Glide Slope
HFEEC	Human Factors Engineering and Ergonomics Center
HITS	Highway In The Sky
HP	High Probability context
HRTF	Head-Related Transfer Function
HR	Heart Rate
HRV	Heart Rate Variability
HVO	High Volume Operation

List of Acronyms (Continued)

IAS	Indicated Air Speed
ICAO	International Civil Aviation Organization
IF	Inspiratory Flow
IFR	Instrument Flight Rules
I-GATE	Integrated General Aviation Training Environment
ISO	International Standards Organization
LCD	Liquid Crystal Display
LLM	Lower Landing Minimum
LLWAS	Low Level Wind Shear Alerting System
LP	Low Probability context
LSA	Loss of Situation Awareness
LTM	Long Term Memory
MCH	Modified Cooper-Harper
MFD	Multi Function Display
MRT	Modified Rhyme Test
MV	Minute Volume
NAS	National Airspace System
NASA-TLX	National Aeronautics & Space Administration - Task Load Index
PAV	Personal Air Vehicle
PC-ATD	Personal Computer – Aviation Training Device
PDC	Pre-Departure Clearance
PF	Pilot Flying
PNF	Pilot Not Flying
PRF	Performance Resource Function
PVI	Pilot Vehicle Interface
RAPID	Rapid Pilot Interface Development Simulator
RIA	Radioimmunoassay
ROA	Roanoke International Airport
RR	Respiration Rate
SA	Situation Awareness
SAGAT	Situation Awareness Global Assessment Technique
SALIANT	Situation Awareness Linked Instances Adapted to Novel Tasks
SART	Situation Awareness Rating Technique
SATS	Small Aircraft Transportation System
SELCAL	Selective Callout
SME	Subject Matter Expert
SUS	Semantically Unpredictable Sentence
SPIN	Speech In Noise
SLM	Sound Level Meter
S/N	Signal to Noise
SO	Station Officer
STM	Short Term Memory
SWAT	Subjective Workload Assessment Technique
SWORD	Subjective Workload Dominance Technique

List of Acronyms (Continued)

TCAS	Traffic Collision Avoidance System
TIS-B	Traffic Information Service - Broadcast
TTS	Text to Speech
UNABLE	Unable to comply
US-DOT	United States Department of Transportation
VFR	Visual Flight Rules
VHF	Very High Frequency
VOR	Very High Frequency Omnidirectional Range
VT	Tidal Volume
WILCO	Will Comply
WL	Workload

VITA

Jeff A. Lancaster

Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University (VT), Blacksburg, VA, 24060-0118
Email: jlancast@vt.edu

EDUCATION

Doctor of Philosophy, Industrial and Systems Engineering, Virginia Polytechnic Institute and State University (Virginia Tech), May 2004. Dissertation: Investigating Pilot Performance Using Mixed-modality Simulated Data Link. Advisor: Dr. John, G. Casali. GPA: 3.7.

Master of Science, Human Factors and Systems, Embry-Riddle Aeronautical University, December 2000. Thesis: Validation of the Boeing Human Modeling System for Use in Air Ambulance Lifting Duty Simulations. Advisor: Dr. John A. Wise. GPA: 4.0.

Bachelor of Science, Biology, Appalachian State University, August 1993. GPA: 3.5.

RESEARCH EXPERIENCE

Doctoral Research, Virginia Tech, 5/2002 – 4/2004.
Explored general aviation (GA) visual flight rules (VFR) rated pilot performance using mixed-modality simulated data link in support of envisioned future US National Airspace System (NAS) contexts. Dr. John G. Casali, Department of Industrial and Systems Engineering.

Masters Research, Embry-Riddle Aeronautical University, 8/1998 – 8/2000.
Investigated the validity of the Boeing Human Modeling System for Air Ambulance Lifting Duty Simulations. Collected, maintained, and analyzed numerous on-site evaluations of air ambulance worker activities. Dr. John A. Wise, Department of Human Factors and Systems.

Research Assistant, Virginia Tech, 8/2001 – 5/2004.
Conducted Hearing Protection Device (HPD) evaluation and testing under ANSI/ISO standards 3.19 and 12.6 A and B. Collected, maintained, and analyzed HPD attenuation results for Noise Reduction Rating (NRR) labeling on consumer products. Drs. John G. Casali and Gary S. Robinson.

Research Assistant, Virginia Tech, 5/2002 – 4/2003.
Designed, collected, and analyzed Human Factors experiments in support of the NASA/Virginia SATSLab Alliance for the Small Aircraft Transportation System (SATS). Drs. Brian M. Kleiner and Gary S. Robinson.

Research Assistant, Embry-Riddle Aeronautical University, 5/1997 – 5/1999.
Designed, collected and analyzed data to support integration of the Computer-aided Performance Analysis System (CAPAS) into the US Navy VS-41 S3-B training squadron at NAS North Island in San Diego, CA. Dr. Jim Blanchard.

Research Assistant, Michigan State University Kalamazoo Center for Medical Studies, 1/1998 – 7/1998. Investigated Human Factors issues related to medical device design and usability. Investigated Human Factors issues of lifting tasks for air ambulance workers using light- to medium-sized helicopters. Dr. John Gosbee.

AFFILIATIONS

- Student Member, Human Factors and Ergonomics Society (HFES), 1997-2004
- Member, HFES Student Chapter, Virginia Tech (VT), 2000-2004
- Member, Alpha Pi Mu Industrial Engineering Honor Society, 2001-2004
- Student Member, National Hearing Conservation Association, 2003-2004.

HONORS & AWARDS

- Dean's List, Appalachian State University, 1991
- Academic Achievement Award, Human Factors and Systems, Embry-Riddle Aeronautical University, 1998, 1999
- United Parcel Service (UPS) Graduate Fellowship for the NASA/Virginia SATSLab Project, Virginia Tech, 2001-2002
- UPS Graduate Fellowship, 2002-2004

GRADUATE COURSEWORK

Human Factors and Systems	Sensation & Perception	Auditory Display Design
Training Systems Design	Human Physical Capabilities	Visual Displays
Human Computer Systems	Memory & Cognition	Usability Engineering
Applied Multivariate Analysis	Optimization	Seminar in Advanced ATC Issues
Aviation Psychology	Macroergonomics	Integrated Systems Design
Human Factors Systems Design I & II	Systems Concepts, Theories, & Tools	HF of Aviation/Aerospace Applications
Human Factors Research Design I & II	Occupational Safety & Hazard Control	Management of Change in Organizational Systems

PUBLICATIONS & PRESENTATIONS

Lancaster, J.A., Olson, S.D., Gardner-Bonneau, D.J., Schoenherr, C., Blostein, P.A. (2001). Lifting problems of air ambulance crews, Ergonomics in Design, 9 (1), Human Factors and Ergonomics Society.

Lancaster, J. A., Saleem, J. J., Kleiner, B. M., Robinson, G. S., and Casali, J. G. (2002). The effect of a six-degree approach angle and lower landing minimum level on pilot performance in a low-fidelity static aircraft simulator (ISE Department Technical Report 200204, Audio Lab Number 9/04/02-4-HP). Blacksburg, VA: Virginia Tech, Grado Department of Industrial and Systems Engineering, Auditory Systems Laboratory. (50 pages) (contractor's report)

Lancaster, J. A., Saleem, J. J., Kleiner, B. M., Robinson, G. S., and Casali, J. G. (2002). Preliminary study on the effects of approach angle and lower landing minimum level on pilot performance in a low-fidelity static aircraft simulator (ISE Department Technical Report 200203, Audio Lab Number 9/04/02-3-HP). Blacksburg, VA: Virginia Tech, Grado Department of Industrial and Systems Engineering, Auditory Systems Laboratory. (77 pages) (contractor's report)

Lancaster, J.A., Saleem, J.J., Robinson, G.A., Kleiner, B.K., Casali, J.G. (2003). Preliminary study on the effect of approach angle and lower landing minimum level on pilot performance in a low-fidelity static aircraft simulator. Human Factors Engineering and Ergonomics Center, Virginia Tech. In Proceedings of the 12th Biennial International Symposium on Aviation Psychology, April 14-17, 2003, Dayton, Ohio USA.

Lancaster, J.A., Robinson, G.S., Casali, J.G. (2003, November). Investigating Pilot Performance Using Mixed-modality Simulated Data Link. Poster presentation at the Fall Meeting of the North Carolina Regional Chapter of the Acoustical Society of America, Winston-Salem, NC, November 7, 2003.

Lancaster, J.A., Robinson, G.S., Casali, J.G. (2004). Comparison of two voice synthesis systems as to speech intelligibility. In Proceedings of the 7th Annual Applied Ergonomics Conference, Orlando, Florida, March 9-11, 2004.

Lancaster, J.A., Robinson, G.S., Casali, J.G. (2004, February). A Human Factors Investigation of Pilot Performance Using Mixed-modality Simulated Data Link. Poster presentation presented at the 2004 National Hearing Conservation Association, Seattle, WA, February 19-21, 2004.

Lancaster, J.A., Robinson, G.A., and Casali, J.G. (2004). Comparison of two voice synthesis systems as to speech intelligibility in aircraft cockpit noise. Accepted to Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting, New Orleans, Louisiana, September 20-24, 2004.