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.

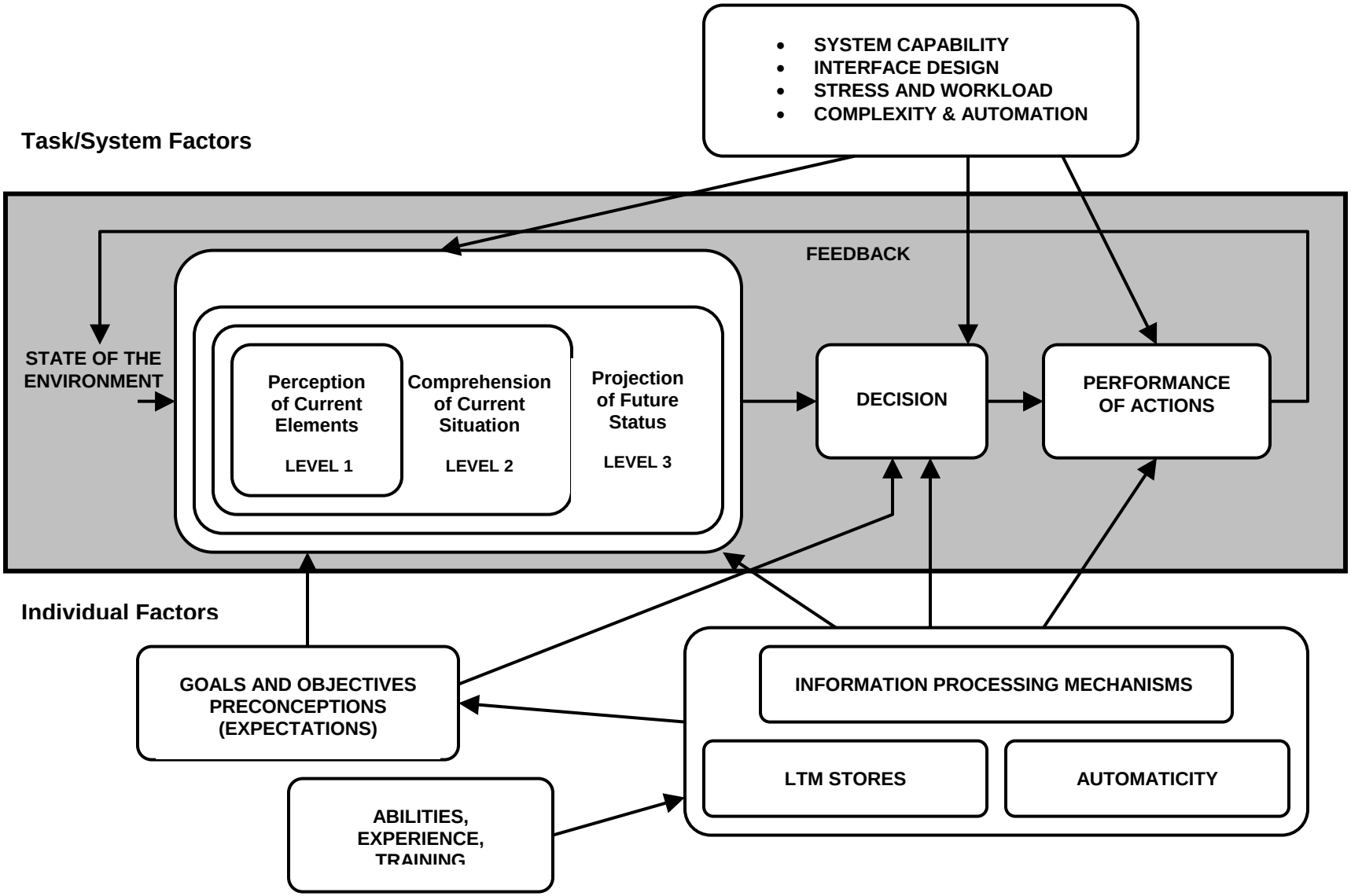

Figure 2. Model of SA in dynamic decision-making. From Endsley & Garland (2000), p. 3.

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

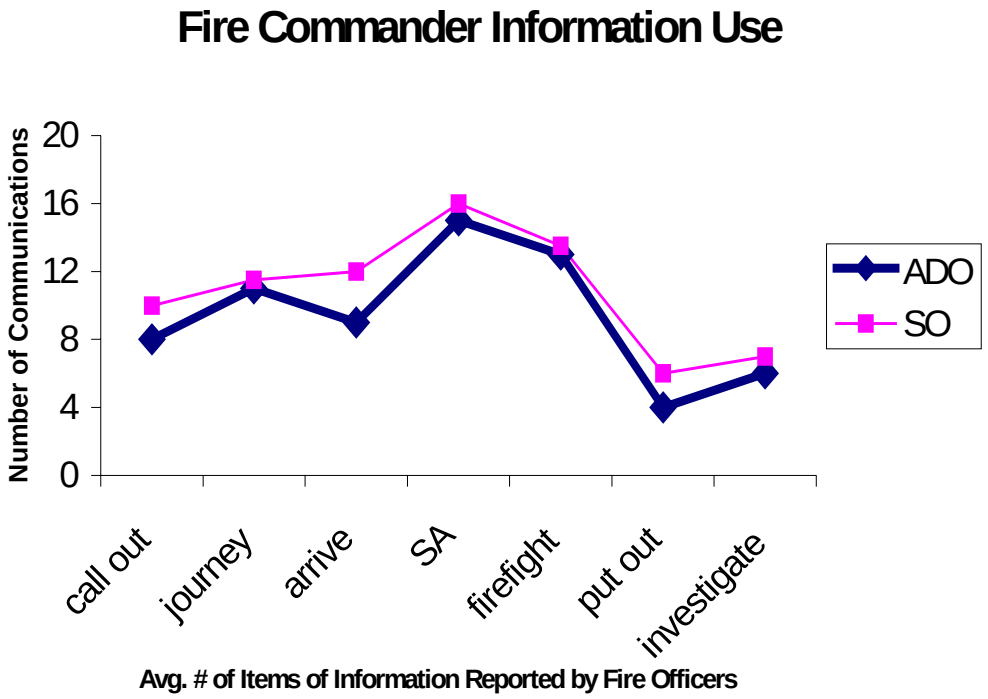

Figure 4. Example of SA requirement method 2 (from Martin and Flin, 1997, p.2/4).

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