THE USE OF SPEECH RECOGNITION TECHNOLOGY IN AUTOMOTIVE APPLICATIONS

Andrew William Gellatly

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

Doctor of Philosophy
in
Industrial and Systems Engineering

Thomas A. Dingus, Chair
Robert C. Williges
John G. Casali
Brian M. Kleiner
Raymond J. Kiefer

March 28, 1997
Blacksburg, Virginia

Keywords: Driver Behavior, Dual-task Performance, Decision Tools

THE USE OF SPEECH RECOGNITION TECHNOLOGY IN AUTOMOTIVE APPLICATIONS

by
Andrew William Gellatly
Committee Chairman: Thomas A. Dingus
Industrial and Systems Engineering

(ABSTRACT)

This dissertation is part of a sponsored research contract, the objectives of which were (1) to perform a detailed review of the existing scientific literature on human factors and speech recognition technology, and the literature supporting the use of speech input for controlling in-vehicle tasks; (2) to develop a decision tool that would aid designers in determining whether a given in-vehicle task function should be performed using speech input or some method of manual control input; (3) to experimentally examine a number of automatic speech recognition (ASR) system parameters, input modalities, and driver ages to evaluate the effects these variables had on driving performance, task-function usability, and driver preference/acceptance of task-function design; and (4) to develop a comprehensive set of human factors guidelines and recommendations for the use of ASR technology in automotive applications. The design guidelines will allow designers and engineers to identify which in-vehicle task functions, when performed while concurrently driving an automobile, are amenable for speech recognition technology, and what ASR system parameters and performance levels are required before implementing ASR technology inside the automobile to control the functions.

The first part of the dissertation is a review of the literature on attentional resource theories, stimulus-central processing-response (S-C-R) compatibility, and mental workload. The literature review suggests that implementing speech recognition technology inside the automobile may improve dual-task performance and reduce mental workload when certain types of in-vehicle tasks are performed concurrently while driving the vehicle. The driving task is primarily a visual-stimulus/manual-response type task. Therefore, concurrent in-vehicle tasks which have auditory-stimulus/verbal-response (or visual-verbal-stimulus/verbal-response) characteristics may be performed best using speech input and interfere least with the primary driving task.

A second part of the literature review examines human factors issues in automatic speech recognition (ASR) technology. The review of scientific literature includes (1) a description of common speech recognition terms and their definitions; (2) the barriers--both technological and behavioral--to implementing ASR technology inside the automobile (e.g., recognition accuracy in noisy environments; system's capability to recognize speaker independent input; etc.); (3) the important ASR system parameters to be evaluated through empirical research; (4) a review of the attentional demands of current in-vehicle tasks; and (5) the implications and potential benefits of incorporating ASR technology to control in-vehicle tasks. This literature review provides some initial guidelines as to what should and should not be included in an automotive speech recognition system for controlling in-vehicle tasks. The review also allowed the number of variables associated with speech recognition systems to be narrowed to a manageable number for empirical evaluation. Only variables that were believed to considerably impact user performance, acceptance, and system usability are included in the experimental methodology section of the proposal.

Next, a decision tool and methodology are described for evaluating whether a given in-vehicle task function should be performed using speech input or some other manual control input method. A series of decision trees were developed as part of the research effort to provide designers and engineers of in-vehicle systems with a systematic

methodology for determining whether a given task function should be performed using manual controls or using speech input. Each decision node and action point on the trees represents various phases of the design decision-making process, leading to the final result of task-function input-modality allocation. Both current and near-future automotive systems can be evaluated for their utility of being controlled via speech input using this decision tool. Decision criteria such as safety, usability, and driver preference/acceptance provide the logical basis for determining whether a given task-function should be considered for control using speech input or some current manual-control input method.

Two experiments were conducted to evaluate the effects of ASR system design parameters, input modality, and age on driving performance, task-function usability, and driver preference/acceptance of task-function design. Each experiment evaluated a combination of ASR system parameters (e.g., recognition errors, error correction methods, and ASR system feedback); or a combination of driver age, input modality, recognition accuracy, and type of recognition error common to the system. Each experiment measured eye movement behavior, steering input behavior, speed maintenance behavior, reaction time to forward scene event, task completion time, and task completion errors of drivers performing various in-vehicle task-functions while concurrently driving. Subjective ratings of driver preference/acceptance as affected by the independent variables were also recorded. The results from the empirical research were used to determine the appropriate ASR system parameters (i.e., system feedback and error correction method) and the appropriate levels of ASR system performance (i.e., recognition accuracy and recognition errors allowed) required before implementing ASR technology inside the automobile to control a variety of task functions.

Results from the two empirical studies showed that manipulation of the various ASR system design parameters (i.e., recognition error type, recognition accuracy, and input condition) significantly affected measures of driving performance, system usability, and driver preference/acceptance. However, from a practical viewpoint, ASR system design parameters had a nominal effect on driving performance. Differences measured in driving performance brought on by changes in ASR system design parameters were small enough that alternative ASR system designs can be considered without impacting driving performance. The speech-input methods tested were not practically different from current manual-input methods used in performing identical in-vehicle tasks for any of the measures of driving performance recorded. Therefore, no benefits can be claimed for ASR systems improving driving safety/performance compared to current manual-control systems.

Automatic speech recognition system design parameters had a moderate impact on the usability of the in-vehicle tasks performed. Criteria such as task-completion times and task completion errors were shown to be different between speech-input and manual-input control methods, and with different ASR system design configurations. Therefore, trade-offs between ASR system designs, and between ASR systems and manual-control systems, can be evaluated in terms of usability. In-vehicle system designs that afford the highest levels of usability should be selected for use in the automotive environment.

Finally, ASR system design parameters were found to have a nominal effect on measures of driver preference/acceptance. Differences found for the subjective opinion data were relatively small and may be of little practical significance. Further research is warranted to determine if long-term use of ASR systems, with less than optimal design parameters, would result in significantly lower driver preference/acceptance compared to the data collected in this research effort.

Human factors recommendations and design guidelines for the use of speech recognition technology in automotive applications are presented in the summary of results and discussion. The recommendations are based on results from the empirical research and the literature review of speech recognition technology and attentional demand of driving.

ACKNOWLEDGMENTS

This research was sponsored by the User Interface Development Group of Delco Electronics Corporation along with support from the Industrial and Systems Engineering Department at Virginia Tech. The following personnel at Delco Electronics are acknowledged for their guidance and insightful comments on this project: Kathy McCombs, Barry Grant, and Dr. Dave Curry. Special thanks go to Kathy and Barry for the encouragement and advice they offered throughout the different phases of the project, and the experience and friendships gained while working as a summer intern with their group.

I would like to thank Drs. T.A. Dingus, R.C. Williges, J.G. Casali, B.M. Kleiner, and R.J. Kiefer for their exceptional patience, time, and contributions to this research effort while serving on my committee. I would especially like to thank the committee chair, Dr. Tom Dingus, for his encouragement and help throughout the tough times of the research. I have learned a lot from Tom and will continue to do so in my current position at the Center for Transportation Research; for that, I am indebted to him. A special thanks goes to Dr. Robert Williges for helping make this project a reality, and for all his encouragement and assistance during the early phases of the research. Thanks also go to Dr. Ray Kiefer for agreeing to serve on my committee and provide valuable input from an industry viewpoint. It was also a pleasure having had the opportunity to work with Ray at General Motors. I would also like to thank Andy Petersen and Brian Daily for their outstanding technical assistance in instrumenting the vehicle and programming the software for my experimental setup. Finally, I want to thank Julie Cook and Shannon Hetrick for their help in data collection, and Marney Davis for her help in data reduction. Without Julie or Shannon, the speech recognition system would have had trouble executing in-vehicle tasks.

I want to thank my parents and family for the love and support they provided during my six-year graduate experience. A special thanks goes to my incredible wife, Liz. She endured the toughest times with me and was there to help every step of the way. Without her love and support, along with that of my parents, I would not have reached my current goals and aspirations.

Finally, I would like to thank all of my friends, fellow graduate students, the faculty and staff at Virginia Tech and the Center for Transportation, and the business associates I have made over the years. All of my experiences and interactions with these people have shaped me and my work, and helped make my graduate experience worthwhile and unforgettable.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The advent of intelligent transportation systems (ITS) has brought new technologies inside the automotive cockpit. Drivers will be able to access greater information than is provided by current instrument panel displays and controls. Navigation, route guidance, traffic management information, collision avoidance, communication systems, and alternative methods for displaying and controlling vehicle information (speed, audio, climate control, engine status, and warning telltales) are just several examples of the new technologies proposed to improve driving performance, comfort, and convenience. Driver interaction with these systems may be different from conventional interface formats. More information will be displayed to the driver, possibly requiring more glances inside the vehicle and for longer durations to assess or control a given system. New systems will compete for space inside the vehicle, forcing displays and controls either to become smaller or to be reconfigurable. Driver interface design will be critical for both user acceptance and performance.

Automatic speech recognition (ASR) technology has been implemented in concept vehicles to demonstrate a new method of interfacing the driver with vehicle control systems. Methods for selecting which in-vehicle systems are to be controlled through speech input are not well defined. Often, systems are arbitrarily selected for speech control due to ease of implementing the speech recognition technology, or the project engineer and/or platform customer believes this is what users should control with speech. Current use of speech recognition technology has been less than optimal in automotive systems from an end-user's viewpoint (see Kobe, 1995).

The goal of this research was to determine the optimal use of speech recognition technology for in-vehicle control of various automotive systems. The research generates methods and guidelines that allow designers and engineers to identify which task functions, when performed while concurrently driving an automobile, are amenable for speech recognition technology. Current and near-future limitations in speech recognition technology are considered in evaluating these tasks. The developed design guidelines are based on functions performed when executing in-vehicle tasks (e.g., changing modes on the climate control system, selecting a specific track from a compact disc in the CD player, etc.) for proposed systems incorporating speech recognition technology. The guidelines consider aspects of system usability, driver performance, driver behavior, and driver acceptance as affected by such ASR system parameters as recognition accuracy, recognition errors and error correction, and system feedback. The methodological approach included the following:

(1) A detailed review of the existing scientific literature on human factors in speech recognition and speech recognition technology.
(2) A decision tree analysis to determine which current and near-future secondary-task functions are amenable to speech recognition technology.
(3) An empirical evaluation of secondary-task functions that are determined both amenable and not amenable to speech recognition technology and the manual analogs of these secondary-task functions.
(4) Development of design guidelines to be used by designers and engineers in determining which automotive control systems should incorporate speech recognition technology.

LITERATURE REVIEW

Automatic speech recognition (ASR) systems may prove useful in situations where the eyes and hands are occupied with control and monitoring functions (Baber, 1991; Bennett, Greenspan, Syrdal, Tschirgi, and Wisowaty, 1989; Cochran, Riley, and Stewart, 1980; Cohen and Mandel, 1982; Cohen and Oviatt, 1994; Jones, Frankish, and Hapeshi, 1992; Markowitz, 1995; Poock, 1986; Quinnell, 1995), when low light or dark levels exist (Poock, 1986), or where speech can provide an additional channel for input to reduce operator workload (Jones et al., 1992; Peckham, 1984). However, ASR systems will only be useful for certain types of functions in complex systems.

Jones, Hapeshi, and Frankish (1989) provide a limited set of qualitative ASR interface guidelines which can be used for the allocation of functions in complex systems. (1) Speech should only be used when input is required infrequently. (2) The assignment of function to modality should preserve the coherence of the task components. Speech should be assigned in a consistent way to one component of the task, for example, use speech for commands only versus commands and data entry. (3) The care with which tasks are assigned to the speech modality should also be extended to functions within the task. For example, speech input is not useful for performing continuous input functions such as positioning a cursor on a display (Murray, Van Praag, and Gilfoil, 1983) or adjusting mirror positions in an automobile (Vail, 1986).

Speech recognition technology may allow the driver of a vehicle to concurrently perform certain in-vehicle secondary tasks without adversely affecting performance of the primary driving task. Currently, in-vehicle secondary tasks are accomplished through manual input from the driver. Automotive systems that are designed to be controlled via manual input may affect driving performance by forcing the driver to look away from the forward road scene and remove one hand from the steering wheel to complete the task. The fact that the driver must divert visual, manual, and/or cognitive attention away from the primary driving task to successfully perform an in-vehicle task raises concerns over how these in-vehicle systems should be designed and what modality of input should be used to control these systems.

The design of automotive secondary systems may also affect driver workload. Measures of driver workload are influenced by the visual demands, the response complexity, and the higher-order processing required by the various in-vehicle displays and controls. Speech recognition technology could reduce the number of hand-operated controls for the various secondary automotive systems. Fewer controls would reduce clutter inside the vehicle and eliminate the need for the driver to select the appropriate control from a large number of available controls. This may result in reduced visual demand and reduced response complexity, which in turn may reduce driver workload. However, the higher-order processing required to remember appropriate command vocabulary and speak the commands may increase driver workload. A well-designed speech interface should make operating secondary systems easier and more convenient than current manual methods. Operator satisfaction while using a system should increase, while the workload imposed by the system should decrease.

Attentional Resource Theories

The ability to divide one's attention between a number of tasks has given rise to the resource metaphor for attention (Wickens, 1992). Currently, there are two theories addressing the resource or capacity of human attention: the Single-Resource Theory and the Multiple-Resource Theory. Both theories view attention as a pool from which all tasks and mental activities are drawn. However, the two theories differ in how they deal with the structural aspects of the tasks being performed (i.e., the input and output modalities used).

The Single-Resource Theory models attention as a single pool of resources. As task demands increase, more resources are drawn from the pool. If the demand for resources exceeds the supply, performance of that task decreases. In addition, if two tasks interfere with each other because they share resources, performance of both tasks decrease.

As an example, driving an automobile demands a certain amount of resources to be drawn from the attentional pool. If traffic density increases or the driver attempts to change the radio station, more resources are required from the attention pool. As long as resource supply exceeds demand, driving task performance should not be degraded. However, if the resources demanded to change radio stations interfere with the driving task resources or the resources demanded to handle increased traffic density exceed the supply, performance of the driving task will decrease. Situations where attentional resource demand exceed supply could result in an accident.

Multiple-Resource Theory argues that numerous capacities exist that have resource properties; this is contrary to the single supply of undifferentiated resources proposed in Single-Resource Theory (Wickens, 1992). Task interference and performance decrements are more likely to occur when common resources are shared. Resources are defined by Wickens (1992) along three dichotomous dimensions: two stage-defined resources (encoding versus responding processes), two modality-defined resources (auditory versus visual encoding), and two processing code resources (spatial versus verbal). Figure 1 shows the structure of multiple resources in a dimensional representation.

By using this model, aspects of dual-task performance and time-sharing may be explained better than by Single-Resource Theory. If two tasks require separate rather than common resources, as defined by the Multiple-Resource model, time-sharing becomes more efficient. Performance of one task will not be as affected if the second task's difficulty increases, provided both tasks do not share common resources.

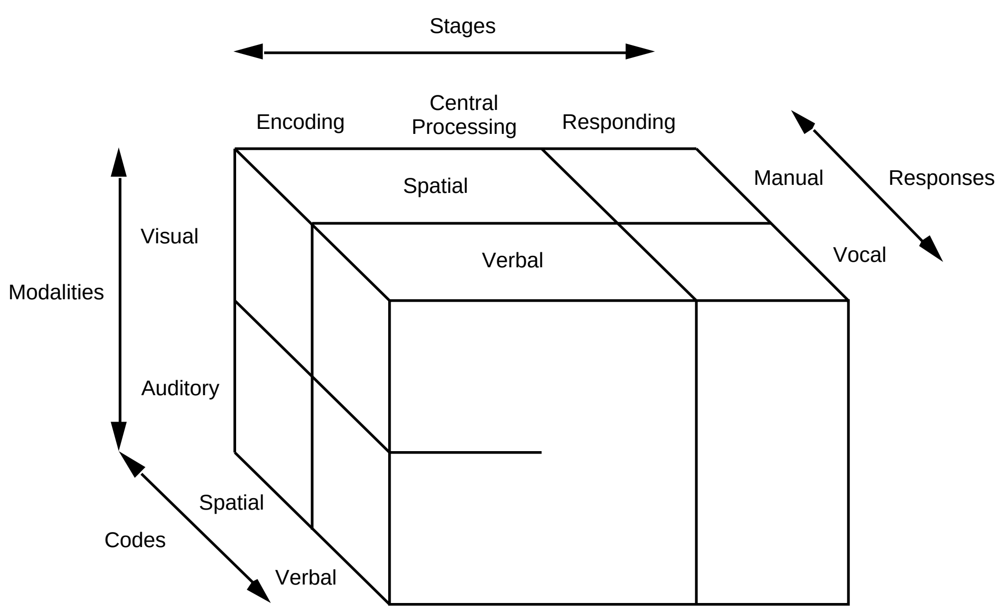

Figure 1. Structure of processing resources proposed by the Multiple-Resource Theory of attention (Adapted from Wickens, 1992).

Revisiting our driving example, Multiple-Resource Theory would predict that primary driving task performance (a visual input with manual response task) would decrease when the driver is concurrently changing the radio station (also a visual input with manual response task). Time-sharing these tasks will be less efficient. However, if the radio provided auditory feedback (e.g., station number currently receiving) and the driver could change radio stations through speech input, Multiple-Resource Theory would predict that primary driving task performance (a visual input with manual response task) would not be affected when the driver is changing radio stations (now an auditory input with vocal response task) and that time-sharing these two tasks will be more efficient.

Trade-offs in task performance and time-sharing result when more tasks are added that demand the same resources. The method in which stimuli are presented to the operator, the way in which the stimuli are processed and coded in the brain, and the method for responding to the stimuli can affect performance in multiple-task situations. An important implication for vehicle operation is to determine what combinations of input, output, and processing are advantageous for driving performance and safety.

Numerous studies have been performed to evaluate the effect of input modality on dual-task performance (Aretz, 1983; Martin, Long, and Broome, 1984; Mountford and North, 1980; Mountford, North, Metz, and Warner, 1982; Tsang, Hartzell, and Rothschild, 1985; Vidulich, 1988; Wickens, Sandry, and Vidulich, 1983; Wickens, Vidulich, and Sandry, 1981). In general, these studies employed a visual-manual primary task (e.g., some kind of tracking task) while having subjects perform concurrent secondary tasks using either manual or voice input to control these tasks. The secondary tasks differed in the modality in which information was presented to the subject either auditorily or visually. Improved dual-task performance was realized when speech controls were used to control secondary task functions when a visual-manual primary task was concurrently time-shared (Aretz, 1983; Martin et al., 1984; Mountford and North, 1980; Mountford et al., 1982; Vidulich, 1988; Wickens et al., 1981). Further, the results demonstrated that best dual-task performance and least interference occurred when the secondary task displayed auditory information--which was verbally encoded--and was responded to with speech (Wickens et al., 1983) or the secondary task displayed visual information--which was verbally encoded--and was responded to with speech (Tsang et al., 1985).

However, the methodologies employed in some of the research on input modality and dual-task performance have been criticized (Baber, 1991; Damos, 1985, 1986). Several of the studies employed a within-subjects experimental design, where the same subjects experienced the different experimental conditions (e.g., both the secondary task speech-input and manual-input conditions). Damos (1985, 1986) and Poulton (1982) report that using a within-subjects experimental design results in asymmetric transfer of learning from one condition to another. Asymmetric transfer means that practice gained by the subject in one experimental condition affects performance in other experimental conditions. Damos (1985, 1986) found no improvement in dual-task performance when a secondary task was performed using speech input versus manual input when a between-subjects experimental design was employed. This led to the conclusion that asymmetric transfer may have occurred in the previous studies on input modality and dual-task performance, even though order for presentation of experimental treatments was counterbalanced. However, key differences in experimental tasks existed between Damos' research (a discrete primary task paired with a discrete secondary task) and the research in question (continuous primary task combined with discrete secondary task). The use of two concurrent discrete tasks in Damos' research could allow subjects to schedule the demands of each task and therefore reduce resource competition (Baber, 1991). Damos' research may have investigated a different aspect of Multiple-Resource Theory instead of competition between processing codes. In addition, some of the criticism discounting the validity of Multiple-Resource Theory was refuted by employing between-subjects experimental designs, replicating previous research, and finding similar results favoring speech input over manual input in improving dual-task performance (Vidulich, 1988).

Stimulus-Central Processing-Response (S-C-R) Compatibility

Driving an automobile imposes a particular load on visual perception and manual response channels. Traffic density or controlling in-vehicle systems (e.g., audio, climate control, navigation) can increase the load on the visual and manual channels. The auditory and speech channels, however, remain largely untapped in current automotive system interfaces. Incorporating the auditory and speech channels into automotive interfaces must be done mindfully. Different display and control requirements arbitrarily assigned to speech and auditory channels may produce a less effective system than if all display and control requirements were assigned to visual and manual channels (Wickens et al., 1983). Stimulus-response (S-R) compatibility influences human information processing and system effectiveness. Therefore, S-R compatibility is an important factor in systems design.

Information presentation modality (visual versus auditory) interacts with response modality (manual versus speech) in that visual-manual and auditory-speech S-R modalities are more compatible and yield greater performance than visual-speech or auditory-manual S-R modalities (Wickens, 1992; Wickens et al., 1981). Tsang et al. (1985), Wickens et al. (1981, 1983), Wickens, Vidulich, and Sandry-Garza (1984), and Vidulich (1988) researched whether these modality-based S-R compatibility relations may also depend on how the task is processed by the brain--verbal versus spatial processing. This body of research employed a manual-visual spatial task as a primary task, and manipulated input, output, and central processing modalities of a secondary task.

In general, the findings from the research by Wickens et al. (1981, 1983, 1984) and Vidulich (1988) demonstrated that dual-task performance and time-sharing were best when the secondary task had verbal central processing and the input-output modalities were in the speech and auditory channels. Since the primary task was a manual-visual (input-output) spatial task, minimal resource competition occurred when the secondary task was a speech-auditory (input-output) verbal task. However, when the secondary task required resources that competed with resources used to perform the primary task (e.g., the secondary task was a manual-visual spatial task), time-sharing efficiency and dual-task performance were significantly degraded. Performance was also found to be sub-optimal when the secondary task was in the speech and auditory channel, but had spatial central processing demands (Wickens et al., 1981, 1984). Interestingly, Tsang et al. (1985) found that a speech-visual (input-output) secondary task resulted in superior dual-task performance over other types of input-output secondary task combinations. This is contrary to the previous findings that a speech-auditory (input-output) secondary task would result in the best dual-task performance because competition for visual resources with the tracking task would be less. Tsang et al. hypothesized that the speech-visual (input-output) secondary task used in their experiment was more compatible than the speech-auditory (input-output) task when performed concurrently with the primary tracking task, therefore resulting in performance that would not be predicted by S-C-R compatibility.

Highest levels of dual-task performance and time-sharing efficiency for multiple-task situations are achieved when speech and auditory input/output modalities are interfaced with verbal tasks (Wickens et al., 1983, 1984; Vidulich, 1988). Automotive secondary-systems can present information either verbally (words and/or numbers) or spatially (icons/symbols or maps). When designing such systems for automobiles, stimulus-central processing-response (S-C-R) compatibility becomes an even more important issue. We must remember that the primary task of driving an automobile requires visual stimulus, spatial central processing, and manual response. An optimal automotive secondary system, therefore, would present auditory stimuli that are verbally processed and responded to through speech. Presenting certain automotive system information auditorily would require some thought to ensure that incompatibilities do not arise accidentally. However, presenting that information visually (along with the other caveats of verbal processing and speech response) would still yield better dual-task performance and time-sharing efficiency than a "visual-verbal-manual" S-C-R-type system.

Mental Workload

The concept of mental workload can be used to evaluate the multi-dimensional aspects of driving a vehicle and performing a concurrent secondary task. Driver workload is assumed in this paper to be comprised of several constructs, including stimulus complexity, higher-order processing, response complexity, and situational awareness. These constructs can be easily quantified when measured individually; however, the effect of each construct becomes less clear when combined under the measurement of workload. When measuring operator workload, the experimenter is attempting to quantify the multi-dimensional aspects affecting task performance into a single measure (workload) which can be used to assess differences in equipment design and/or operator skill level. In other words, workload measurement attempts to quantify the attentional demand and S-C-R compatibility when performing tasks into a single measure which can be used to assess multiple-task performance (Wickens, 1992). Therefore, workload should be an important concern in automotive applications, particularly when considering implementation of secondary systems that may affect driver workload in a negative way. Increasing workload on the driver can amplify resource competition and S-C-R compatibility effects (Wickens et al., 1983).

Mental workload is loosely defined as the ratio of resources demanded to perform a task to the resources a person has available (Sanders and McCormick, 1987; Wickens, 1992). Workload is a measurable quantity that can evaluate the demands a particular task or group of tasks exert on a driver's information processing. The amount of workload a driver experiences while interacting with a system can change as a result of changes in resources demanded by the task (e.g., the task must now be performed under low illumination conditions) or changes in available operator resources (e.g., the amount of mental resources available may decrease because the driver is fatigued). To account for changes in workload, the driver can either allocate more resources to the task (provided there are resources still available), or task demands can be altered to incorporate additional information-processing channels.

Speech input systems may be useful for reducing operator workload, especially when the operator must perform two tasks concurrently, such as fly an airplane and enter navigation waypoints, or drive an automobile and select a destination from a navigation system (Aretz, 1983; Jones et al., 1992; Mountford and North, 1980; Mountford et al., 1982; Mountford, Schwartz, and Graffunder, 1983). An ASR system can provide an additional response channel over which the workload on the operator can be distributed (Martin, 1989).

The assessment of workload imposed by equipment is an important technique for system designers (Sanders and McCormick, 1987; Wickens, 1992). Measures of workload are broadly classified into four categories: primary-task measures, secondary-task measures, physiological measures, and subjective measures. Subjective measures of workload typically use rating scales that elicit either a one-dimensional or multidimensional assessment of information processing load. All measures of workload should meet the following criteria (Sanders and McCormick, 1987; Wickens, 1992):

(1) *Sensitivity*. The workload measure should distinguish between tasks requiring more resources and those requiring less, and be sensitive to changes in task difficulty.
(2) *Selectivity*. The workload measure should only be influenced by differences in resource demand, not extraneous factors such as physical workload or emotional stress, which may not be related to information processing.
(3) *Diagnosticity*. The workload measure should not only indicate changes in workload, but also what may have caused the change (e.g., is the change attributable to changes in resource demand or information channel processing).
(4) *Obtrusiveness*. The workload measure should in no way interfere with, disrupt, or contaminate performance of the task whose workload is being assessed.
(5) *Reliability*. The workload measure must reliably measure information processing demands, and also be sensitive to rapid changes in demands that may occur during the measurement.

A potential limitation for the assessment of driver workload is that many of the methods used for measuring workload may not be sensitive enough to pick up differences in workload between performing an in-vehicle secondary task using conventional manual input and performing the same task using speech input. Differences in visual demand (stimulus complexity) and manual demand (response complexity) are expected to exist between ASR-controlled tasks and similar manual-controlled tasks. Visual demand and manual demand are two constructs that comprise workload. If differences in workload between the two input methods (manual versus speech) are to be found, the measurement technique chosen to evaluate driver workload must be sensitive to these differences. It is important to note that the measurement technique chosen must also be sensitive to changes in the other constructs that comprise driver workload (i.e., higher-order processing and situational awareness). Even though differences in visual and manual demands are expected between the two input methods, the cognitive demands (higher-order processing) imposed by the speech input method may increase to a point where no difference in overall driver workload is found. Whichever measurement technique is chosen should demonstrate diagnosticity. Because diagnosticity is a difficult achievement for most driver-workload measurement techniques, it is recommended that other measures (e.g., micro-measures of workload such as eye-movement behavior, task-completion time, response-time to forward scene events, etc.) be collected when evaluating driving performance to determine differences in resource demand, S-C-R compatibility, and task difficulty when performing various in-vehicle secondary tasks.

Speech Recognition Terms and Definitions

The purpose of this research is to determine when, where, and how speech recognition technology should be used in the driver-automobile interface. Guidelines for applying speech recognition technology will depend on many variables, some of which include characteristics of potential users, the physical environment, the driver's workload, the constraints imposed by the task, and the ASR system itself (Simpson, McCauley, Roland, Ruth, and Williges, 1985, 1987). The following section contains a list of terms and definitions that are important to automatic speech recognition technology.

Speaker dependence. One of the parameters that differentiates ASR systems is the extent of data regarding voice characteristics of potential users required by the system to perform effectively. Speech recognition systems that require specific examples of a user's speech are defined as speaker-dependent systems (Baber, 1991; Cohen and Oviatt, 1994; Makhoul and Schwartz, 1994; Markowitz, 1995; Pallet, 1986; Pelton, 1993; Poock, 1986;

Quinnell, 1995; Schafer, 1994; Simpson et al., 1985, 1987; Wattenbarger, Garberg, Halpern, and Lively, 1993). These systems must have a sample of how a user will say each vocabulary item before the system can recognize the user speaking the item (Fink, 1981; Lee, Hauptman, and Rudnicky, 1990). Speaker-dependent systems, in general, demonstrate higher recognition accuracy than other types of ASR systems because the system is calibrated for a specific user's voice (Bristow, 1986).

Speech recognition systems that are capable of recognizing speech spoken by any potential user in a given language are defined as speaker-independent systems (Baber, 1991; Cohen and Oviatt, 1994; Fink, 1981; Makhoul and Schwartz, 1994; Pallet, 1986; Pelton, 1993; Poock, 1986; Quinnell, 1995; Schafer, 1994; Simpson et al., 1985, 1987; Wattenbarger et al., 1993). These systems do not require examples of speech from every potential user. Instead, the system uses algorithms that are robust enough to recognize anyone who talks to it in a specific language. Speaker-independent systems can achieve similar levels of recognition accuracy as speaker-dependent systems; however, this is usually true for only small vocabulary sets of less than 50 acoustically-distinct utterances (Bennett et al., 1989; Bristow, 1986; Doddington and Schalk, 1981; Quinnell, 1995). Speaker-independent systems are useful when the user population is large or when multiple users may access the application infrequently (Markowitz, 1995).

Speech recognition systems that begin as speaker-independent ASR systems but are calibrated to the speech of individual users as the system receives samples of users' speech are defined as speaker-adaptive systems (Baber, 1991; Markowitz, 1995; Quinnell, 1995). These types of ASR systems are designed to be used by a stable set of operators who will access the system repeatedly over time. Recognition accuracy is quite high for these systems because voice characteristic samples for each vocabulary item are continuously updated each time the user interacts with the system.

Speech variability. Five levels of variability in spoken language are described by Simpson et al. (1985, 1987) that pose challenges for developers of speech recognition systems. These levels of speech variability include language families, individual languages, dialects, idiolects, and variations in an individual’s speech over time or due to stress, fatigue, or state of physical health. Examples of these speech variables are shown in Table 1. The two levels of speech variability that impact ASR system performance the most are the idiolect and idiolect variation. An idiolect is defined as the language or speech pattern of a single individual. An individual's speech pattern can vary over time and as a result of stress, fatigue, illness, or emotion (Costet, 1982; Markowitz, 1995; Pallett, 1986; Pelton, 1993; Quinnell, 1995; Roe and Wilpon, 1993).

Speaker-dependent ASR systems are better able to account for the variability in speech patterns between individuals because the system must first have a sample of how every user will say each vocabulary item. Speaker-independent ASR systems grossly simplify the breadth of speaker variability. Therefore, speaker-independent systems are usually less accurate in recognizing spoken vocabulary items from individuals whose idiolect or idiolect variation were not considered when the speaker-independent algorithms were developed. The advent of speaker-adaptive ASR systems has resolved some of the challenges associated with human speech variability.

Table 1. Levels of variability in spoken language (Adapted from Simpson et al., 1987).

Linguistic Level	Example
Language family	Indo-European
Individual language	English
Dialect	American English
Idiolect	Individual male person
Idiolect variation	Individual speaking February 9, 1996, while addressing a large lecture hall full of students

Speaking mode of user. A third parameter that differentiates ASR systems is the manner in which a user must speak to the system. Speech recognition systems that require the user to pause briefly (approximately 100 to 200 msecs.) between each spoken vocabulary item are referred to as isolated-word recognizers (Beek, Neuberg, and Hodge, 1977; Cohen and Oviatt, 1994; Lea, 1980a, 1980b; Martin, 1976; Martin and Welch, 1980; Pelton, 1993; Poock, 1986; Rabiner, 1995; Simpson et al., 1985, 1987). These systems may also be referred to as discrete-input speech recognizers (Fink, 1981; Lee et al., 1990; Markowitz, 1995). Recognition accuracy is increased for isolated-word recognizers because the beginning and ending of each word are clearly defined.

Speech recognizers that do not require artificial pauses between each vocabulary item are defined as connected-word recognition systems (Fink, 1981; Lea, 1980a; Lee et al., 1990; Markowitz, 1995; Poock, 1986; Rabiner, 1995; Simpson et al., 1985, 1987). Connected-word recognizers require the user to speak with the same intonation pattern that would be used if the vocabulary items were read from a list. A distinction is usually made between connected-word and continuous-speech ASR systems. A continuous-speech recognition system allows the user to speak in his or her natural speech rhythm and intonation without having to pause between each vocabulary word (Cohen and Oviatt, 1994; Lea, 1980a, 1980b; Martin and Welch, 1980; Pelton, 1993; Poock, 1986; Rabiner, 1995; Simpson et al., 1985, 1987).

Finally, some ASR systems are capable of continuous speech-understanding (Lea, 1980a; Lee et al., 1990; Simpson et al., 1985, 1987). Such systems attempt to correctly accomplish tasks using continuous speech input. The measure of performance of such systems is not word recognition or message recognition accuracy, but response accuracy. These systems attempt to assign meaning to the messages that are spoken to them by the user and respond accordingly.

It is important to note that as the speaking mode of an ASR system changes from a simple isolated-word recognizer to a continuous-speech recognizer, the computer processing power required to accurately perform such recognition actions also increases. Computer processing power has a cost. Therefore, it is important to understand the task or tasks performed by the user so that an appropriate level of speaking mode (and computer processing power) can be selected.

Vocabulary size. A fourth parameter that differentiates ASR systems is the type and size of the vocabulary of the system. In general, there are two types of vocabularies--fixed (or limited) and unlimited (Pallet, 1986; Pelton, 1993; Simpson et al., 1985, 1987). Fixed-vocabulary ASR systems require human speech examples for each vocabulary item that the system is capable of recognizing. Speech recognition applications are designed by extracting vocabulary items from dictionaries of possible speech commands. Acoustical pattern matching is performed between the user's spoken word and words stored in the dictionaries (Markowitz, 1995; Simpson et al., 1987). The size of these vocabulary dictionaries can range from a dozen words to more than 100,000 words. The vocabulary dictionaries are built into the ASR system by the designers and can be modified using a speaker-dependent- or speaker-adaptive-type strategy in ASR system design.

An unlimited-vocabulary ASR system does not require a built-in dictionary of words that are spoken by potential users of the system. Instead, unlimited-vocabulary systems utilize algorithms that analyze spoken utterances into phonetic segments (Simpson et al., 1985, 1987). A phoneme is the smallest speech sound that can change the meaning of a word. Once the phonetic segments are analyzed, the system tries to determine which word was actually spoken and generate an appropriate response. Unlimited-vocabulary systems are useful for applications such as dictation products where the number of possible words spoken by the user can be infinitely large.

Enrollment. Another parameter associated with ASR systems is the type of enrollment a particular system is capable of allowing. Enrollment is the process of providing speech examples (templates) to the system for each of the vocabulary items found in the system's dictionary (Bennett et al., 1989; Cohen and Mandel, 1982; Pallett, 1986; Simpson et al., 1985, 1987). A human speech waveform is converted into a digital template for each vocabulary item. When the user interacts with the ASR system, the speech waveform is converted to a digital template, and the system tries to match that template to the most similar matching template stored in the system's memory.

User enrollment is found on speaker-dependent ASR systems. Each user of a speaker-dependent system must be enrolled if an appropriate level of recognition accuracy is to be achieved. Speaker-independent systems are designed to be enrolled by the vendor. Vendors of a speaker-independent system develop templates that will result in the best recognition accuracy for a large population of potential users. Speaker-adaptive ASR systems start out as vendor-enrolled systems, allowing large numbers of users to use the system immediately, but these systems are capable of user enrollment, which improves the recognition accuracy as the user interacts with the system.

Grammar. Search time to find an appropriate match to a spoken utterance increases as the size of the vocabulary increases. Many ASR systems are beginning to employ grammars to reduce the search time by specifying what words can occur at any point in the interaction with the system (Costet, 1982; Lee et al., 1990; Makhoul and Schwartz, 1994; Markowitz, 1995). An example of a grammar for an automotive ASR system may allow the following input. A driver wants to play a cassette tape already in the tape player of the audio system. The driver may say the following commands: "radio" (pause), "power" (pause), "tape" (pause), "play." The input is defined by first selecting a secondary automotive system from the available choices: radio, climate, phone, and trip. The second and third inputs are selections from the available choices under radio commands: power, AM/FM, tape, CD/Disc, mute, preset, scan, and cancel. The final input is a selection from the available choices under tape commands: stop, play, next, previous, fast forward, rewind, reverse, and cancel. In this example alone, there are 20 vocabulary items (not including vocabulary used to control the other automotive systems or functions within each system). At any one level in this example, the ASR system only has to search from as few as four to as many as eight possible matches rather than search all 20 for a possible match.

Grammars not only reduce search time, but also increase accuracy by reducing the probability of similar sounding words existing in the same list of potential matches at any given point during the interaction (Quinnell, 1995). Words that sound similar but have different meanings are called homonyms (e.g., the number "four" and the preposition "for"). It is the job of the ASR system developer to ensure that only acoustically-distinct vocabulary items exist at each level of potential matches so that recognition speed and accuracy are optimized.

Three types of grammars have been described in the literature: finite-state grammar, statistical models, and key-word spotting (Brems, Rabin, and Waggett, 1995; Markowitz, 1995; Quinnell, 1995). Finite-state grammars are similar to the example above where only a small subset of vocabulary items--a word list--are available at each stage of the interaction. Statistical models are used in larger vocabulary systems and utilize word-sequencing probabilities to determine appropriate matches. These probabilities are derived from analyzing the language samples used to interact with the ASR system. A dictation product, for example, may contain statistical models that determine the probabilities that a spoken utterance follows a previously correctly-recognized word. The word "for" may have a lower probability of following the recognized word "number" compared to its homonym "four," depending on the application and the expected dialogue between the user and the ASR system.

Key-word spotting is used in applications where the user is unaware of the exact wording required by the application. Key-word-spotting ASR systems scan the continuous-speech input for embedded words that are recognizable by the system (Beek et al., 1977; Brems et al., 1995; Juang, Perdue, and Thomson, 1995; Lea, 1980a, 1980b; Markowitz, 1995). For example, a driver wants to use an in-vehicle navigation system to find a place to eat dinner. The ASR-controlled navigation system is capable of key-word spotting. The navigation system is looking for the keywords "food," "restaurant," or "eat," and would display the appropriate destinations from any of the following inputs. "I'm hungry, where can I get something to eat?" "What kind of restaurants do they have in this town?" "I want to see a selection of fast-food places nearby."

Barriers to Implementing ASR Technology

Automatic speech recognition technology has progressed tremendously over the past two decades. However, the technology is still years away, if ever, from being as capable at recognizing and understanding speech as its human developers. Designers of complex systems must exercise care before implementing ASR technology as a method of interfacing an operator with a machine. The tasks to be performed by the operator, the criticality of task execution, the environment in which tasks are to be performed, and the characteristics of the user population all must be understood before implementation of ASR technology occurs. By understanding the nature of the tasks to be performed and the users' characteristics, appropriate tradeoffs in ASR system performance parameters can be made which will result in optimizing the overall usability and acceptance of this person-machine interface.

Technological barriers. Despite the tremendous gains in ASR technology over the past two decades, several parameters associated with ASR system performance still exist that can negatively influence ease of use and user acceptance of products incorporating such technology. Near-future advances in microprocessor technology, digital signal processing, and acoustics may overcome some ASR system performance limitations, such as recognition speed and environmental factors. However, most of the performance parameters described in this section are viewed as long-term technological barriers for the implementation of ASR technology in unique environments (such as an automobile).

Recognition accuracy describes the frequency that a user's spoken utterance is correctly recognized when the spoken utterance is a valid audio input (Casali, Williges, and Dryden, 1990; Fink, 1981). Recognition accuracy is the most commonly used measure of performance for ASR systems, and is viewed as one of the main limitations of this input technology (Simpson et al., 1985, 1987). There are many factors that influence recognition accuracy, such as speaker dependence, speaking mode of the user, speech variability between users, ambient noise level, vocabulary size, acoustic similarity of vocabulary words, and length of spoken word. Some manufacturers of ASR systems claim levels of accuracy that approach 100%. However, these accuracy levels are measured under controlled laboratory conditions. Such high levels of recognition accuracy are often difficult to obtain in the application's true operating environment. Tradeoffs must be made between ASR system features and the desired level of accuracy for a given application.

A second parameter, related to recognition accuracy, is discrimination accuracy. Discrimination accuracy describes the frequency that an invalid auditory input is rejected (Fink, 1981). Designers of ASR applications try to achieve high levels of discrimination accuracy so that non-vocabulary items spoken by the user, or extraneous noises in the operating environment, are not recognized as valid commands by the system. Insertion errors (described in detail below) result when a non-vocabulary word or sound is recognized by an ASR system. It is more natural in human conversational speech to repeat a word that was not understood (or recognized) than to have to correct the listener for a word that was misinterpreted (or misrecognized). High levels of discrimination accuracy can result in lower levels of "first-time" recognition accuracy because the system is biased to reject inputs that do not have a good template match. Again, tradeoffs must be made between desired recognition accuracy and the types of errors that may result from lower levels of discrimination accuracy. In general, it is preferred to have the user repeat a vocabulary item that is rejected than to have the user correct an error by the system (Fink, 1981).

Three mutually exclusive categories of recognition errors are described in the literature (Baber, 1991; Bekker, Van Nes, and Juola, 1995; Cohen and Mandel, 1982; Fink, 1981; Jones et al., 1992; Pelton, 1993; Simpson et al., 1985, 1987). The first category of recognition errors is substitution errors. A substitution error is recorded when a spoken vocabulary item is mistaken for another vocabulary item stored in the system. An example of this type of error would be the user saying the vocabulary item "tape" and the ASR system recognizing the spoken word as another vocabulary word "disc." The second category of recognition errors is rejection errors. A rejection error describes when a spoken word that is a valid vocabulary item is detected by the ASR system, but is not recognized. A user may say the valid vocabulary word "climate," but the system fails to respond in any fashion. The third category of recognition errors is insertion errors. An insertion error occurs when a vocabulary item is recognized but the user did not speak a valid vocabulary word. An example of this would be the ASR system recognizing the word "phone" while the user was having a conversation with another person, but the word "phone" was never spoken. Substitution errors and insertion errors are least tolerated by users because they require the user to correct the input error, which typically necessitates more steps than simply repeating a vocabulary word that was initially rejected by the ASR system.

Recognition speed describes the amount of time that elapses between the end of the user's spoken utterance and the user's ability to perceive that the machine has reacted to the utterance (Fink, 1981; Peckham, 1986). Like the other parameters associated with ASR system performance, recognition speed is influenced by other factors. Recognition speed is directly related to the size of the active vocabulary. The more words a system has available to compare to a spoken utterance, the longer the duration is to find an appropriate match. A faster microprocessor can reduce this search time (Lea, 1980b). However, the best method for increasing recognition speed (or response time) is to limit the available vocabulary that is searched by employing one of the grammar types described previously (Kamm, 1994; Lee et al., 1990; Peckham, 1986).

A related parameter to recognition errors is the error correction method allowed by the ASR system. It is important to provide a convenient and reliable mechanism for correcting errors if user acceptance of speech input is to remain at high levels (Martin and Welch, 1980). The method of error correction used depends on the specific application and the type of feedback provided by the ASR system. One method for correcting recognition errors is for the user to say "yes" or "no" after each vocabulary item is recognized and displayed. A "no" response by the user deletes the last recognized vocabulary item and the user repeats the misrecognized word. This method becomes extremely time consuming and irritating to the user for certain applications, such as a dictation product. A second method of error correction is to have the user respond only to misrecognitions by saying a command word such as "cancel" or "delete." Correct responses by the ASR system would not require a response from the user. This method can be extended to entire sentences or command phrases. A buffer would store vocabulary items as they are spoken, and this buffer could be edited before the commands are executed by the system (Baber, 1991; Martin and Welch, 1980). An appropriate type of feedback is required by systems using a buffer, and the editing commands must be carefully selected so that misrecognition of the editing-command vocabulary does not occur. A major drawback with buffer and command-editing systems is that the use of an error correction dialogue adds an intervening task between the user and the primary task (e.g., calling home from a hands-free cellular phone). A third method of error correction is to allow the user to repeat a command word or string, after which the system determines which vocabulary items need correction. A certain degree of "intelligence" is required by the system to determine which vocabulary words are being corrected. A final method of error correction automatically detects illegal input sequences and corrects them to the most likely legal sequence (Simpson et al., 1985, 1987). The ASR system can either automatically correct the recognition error without alerting the user, or can present the user with a "next-best-alternative" to the misrecognized word. Presenting the user with an alternative choice places control of error correction back to the user. Spelling checkers found on most word-processing programs are an example of this fourth method of error correction. Some word-processing programs automatically correct users' common spelling mistakes. These same programs also provide a list of alternative words to select from when a misspelled word is discovered. A balance must exist between fluent interaction and the inconvenience of recovering from recognition errors (Jones et al., 1992). Selecting the appropriate method of error correction for a given application can lead to this balance and increase user satisfaction with an ASR system.

Users of ASR systems must be provided with some type of feedback to correct recognition errors (Schurick, Maynard, and Williges, 1983). Appropriate feedback can also enhance ASR system performance. In general, two levels of feedback and two modalities in which feedback can be presented are described in the literature. The two levels of feedback are termed primary and secondary (Baber, 1991; Baber, Usher, Stammers, and Taylor, 1992; Jones et al., 1989, 1992; Peckham, 1986). Primary feedback occurs when the system responds immediately to a recognized command or vocabulary item. For example, the driver hears the compact disc begin to play after speaking the command "play disc," or the system responds with a message that the climate control mode has been changed to "max A/C." Secondary feedback occurs when vocabulary items are repeated back to the user for verification before any action is taken by the system. A cellular telephone may repeat each digit spoken by the user and initiate the call only after the user has given the command "send." Any misrecognition of a spoken digit can be corrected before the call is initiated. Primary feedback is preferred in situations where errors are not critical or task completion time must be minimized (Peckham, 1986). Secondary feedback is preferred when accuracy is important, either because the action is safety-critical, or because recovery from an error is difficult (Jones et al., 1992).

The modalities in which feedback can be presented are via the auditory channel, the visual channel, or both (Baber, 1991; Baber et al., 1992; Costet, 1982; Forren and Mitchell, 1986; Isenberg, Yntena, and Wiesen, 1984; Jones et al., 1989; Martin, 1976; Martin and Welch, 1980; Schurick et al., 1983). The best modality for providing feedback depends on the tasks involved and the environment in which tasks are performed. Steiner, Burkhart, and Berson (1983) found that visual feedback interfered with performance of the primary task when subjects concurrently controlled a primary visual-task while interacting with an ASR system. They found that the combination of both auditory and visual feedback interfered least with performance of the primary task and provided the greatest likelihood that subjects would attend to the feedback. Driving an automobile is primarily a visual task. Therefore, either auditory or auditory plus visual feedback should result in the least degradation to driving performance when drivers are concurrently interacting with an ASR system. Whichever level and modality of feedback is selected, the feedback message should not delay the user in completing the task (Cosky, Lively, Roberts, and Wattenbarger, 1995).

Brems et al. (1995) experimentally tested a natural language reprompting strategy for conditions when an ASR system was unable to interpret a spoken command. When an ASR system is unable to understand a spoken command even though a valid command has been spoken, and the system performs no action, it is the same as if a rejection error had occurred. Brems et al. hypothesized that a natural language reprompting, or feedback, strategy (e.g., "Sorry, please repeat" instead of the traditional "Your request was not understood" feedback) would be more appropriate for correcting rejection errors by the system. Users are more likely to repeat what they just said if a natural language reprompt is used instead of no response by the system or a typical error message response such as "did not understand." The ASR system used was modeled after an automated telephone-operator service. Subjects were instructed to perform a series of automated operator-assisted telephone calls such as collect, calling card, third number, and person-to-person calls. Results indicated significantly faster telephone-task transaction times using natural language feedback to reprompt users versus traditional ASR system feedback (i.e., "Your request was not understood"). Therefore, a similar natural language feedback message may improve in-vehicle task performance and increase user preference/acceptance for situations when a rejection error occurs during an interaction with in-vehicle ASR systems.

Environmental factors such as ambient noise, room acoustics, and transmission channel can affect the speech input to an ASR system, resulting in degraded performance (Lea, 1980b; Pallett, 1986; Roe and Wilpon, 1993). The use of noise-canceling microphones and engineering methods to reduce ambient noise and improve room acoustics can improve recognition accuracy under adverse operating conditions. In addition, various kinds of signal-processing methods may be employed by an ASR system to suppress additive noises and improve system performance (Atal, 1994; Furui, 1994). A common method for improving system performance is enrolling speaker-dependent ASR systems in the actual environment in which they are to be used. By enrolling the system in its operating environment, ambient noise conditions and room acoustics are included as part of the speech template that is used in matching future inputs. The speech-signal to noise recorded as a speech-template will provide a better match to actual speech inputs than if the speech-template was recorded in the absence of any ambient noise. The environment inside an automobile can have high levels of ambient noise (>70 dBA sound pressure level) which must be considered when developing an ASR system to control secondary automotive systems.

Behavioral barriers. User characteristics can significantly impact ASR system performance. This section describes several parameters related to the user of an ASR system that are viewed as behavioral barriers for the implementation of ASR technology in certain environments, such as an automobile.

The speaking mode of users can change over time because of fatigue or changes in levels of stress or workload. It is unnatural for people to speak with brief pauses between each spoken word, as is required by isolated-word ASR systems. As the user becomes fatigued or levels in workload change, the user will revert to his or her natural "connected" flow for speech. Recognition errors can result when an ASR system requires isolated-word inputs but the speaking mode of the user has changed to continuous speech. For example, the phrase "wreck a nice beach" when spoken in a continuous speech flow may be recognized by an isolated-word ASR system as the phrase "recognize speech." Therefore, speech recognition systems that handle continuous speech input should be selected for operating environments where the speaking mode of the user may change and impact task performance.

The operating environment can affect the speech variability of the user, which negatively impacts ASR performance. Physical stress, such as vibration, heat, or acceleration, increases the variability of a user's speech input to an ASR system. Psychological stress, such as anxiety, fatigue, or workload, can also increase speech variability. The speech recognition system selected for use must account for the variability in speech that may be found in the given operating environment. Speaker-adaptive systems are able to maintain high levels of performance even when a user's speech varies over the course of interacting with the system.

User acceptance of an ASR system is influenced by all the parameters described under the "technological barriers" section and some additional parameters, such as speaker dependence, speaking mode, vocabulary size, enrollment, and grammar. Recognition accuracy is a primary determinant of user acceptance. If an ASR system could recognize speech input with 100% recognition accuracy, other variables, such as speech variability, enrollment, recognition errors, error correction, feedback, and environmental factors, would become inconsequential. The selection and use of speech recognition technology is highly dependent on the tasks to be performed by the user, and the operating environment in which the technology will be used. Certain tasks may be performed equally as well or better using a manual input technique without having to consider all the variables associated with speech recognition technology. The system designer must be aware of the tradeoffs to make correct decisions about the person-machine interface.

Important ASR System Variables for Experimental Consideration

This section addresses the variables that are perceived to be of greatest importance when considering an ASR system's impact on task performance and user acceptance. The determination of which variables affect task performance, ease of use, and acceptability is based heavily on recommendations from the literature and, when available, experimental data. It is important to note that these variables are highly interdependent. Changing the level of one parameter may affect another parameter's influence on task performance and/or user acceptance. Furthermore, the type of tasks being performed, the environment in which tasks are performed, and the users performing the task play a significant role in determining what levels of each variable are selected so that acceptable levels of performance and user satisfaction are realized.

Recognition accuracy. This variable is probably the most important ASR system parameter to consider when evaluating the utility of speech recognition in performing tasks. Not all applications incorporating ASR technology require 100% recognition accuracy. Applications where recognition errors can be tolerated, interaction with the ASR system is infrequent, and/or task completion time is not critical may be suitable for low-cost speech recognition systems which perform at lower levels of recognition accuracy. The tasks to be performed using an ASR system must be fully understood, as well as the characteristics of the potential users of the system. Once the tasks and user population are defined, intelligent decisions about ASR parameters such as recognition accuracy can be made. The following references discuss the importance of recognition accuracy, how it is affected by other system parameters, and how it may affect task performance and ease of use:

- Baber, 1991
- Bernold, 1993
- Brandt, 1987
- Casali et al., 1990
- Cohen and Mandel, 1982
- Doddington, 1980
- Doddington and Schalk, 1981
- Fink, 1981
- Juang et al., 1995
- Lea, 1980b
- Lee et al., 1990
- Makhoul and Schwartz, 1994
- Markowitz, 1995
- Martin, 1976
- Pallett, 1986
- Peckham, 1984
- Quinnell, 1995
- Roe and Wilpon, 1993
- Simpson et al., 1985, 1987
- Wattenbarger et al., 1993

Only one reference experimentally manipulated the variable "recognition accuracy" to determine its role in system performance and user acceptance (Casali et al., 1990). In general, their results show that as recognition accuracy increases, performance while interacting with the system and user acceptance also increase.

Recognition speed. The speed with which an ASR system responds to a user's speech input can influence ease of use and system acceptance. However, recognition speed is really only critical for tasks that are time-dependent (e.g., a task must be executed in less than one second after a command has been spoken). Vocabulary size has the greatest influence on recognition speed because the ASR system attempts to make an appropriate match to the spoken input from the available vocabulary. The use of appropriate grammars and a fast microprocessor can reduce or eliminate the problem of slow response to speech input. The following references discuss the importance of recognition speed, how it is affected by other system parameters, and how it may affect task performance and ease of use:

- Baber, 1991
- Fink, 1981

- Jones et al., 1989
- Kamm, 1994
- Lea, 1980b
- Lee et al., 1990
- Newell, Arnott, Carter, and Cruickshank, 1990
- Peckham, 1986

No references were found that experimentally manipulated the variable "recognition speed" to determine its role in system performance and user acceptance.

Recognition errors and error correction. The types of recognition errors made by an ASR system and the available method to correct errors can affect task performance and user acceptance. Recognition errors such as substitution errors or insertion errors require the user to perform additional steps to correct the error. If these types of errors are frequent and the method to correct the error is time consuming or difficult to perform, the user becomes dissatisfied with the interaction method, and task performance suffers. Again, other speech recognition parameters affect what types of recognition errors occur and what error correction method is best for a given application. In general, substitution errors can be reduced by selecting acoustically-distinct command vocabulary, and insertion errors can be minimized by incorporating a noise-canceling microphone and a "push-to-talk" switch to be used when speaking commands to the ASR system. The following references discuss recognition errors and error correction, how they are affected by other system parameters, and how task performance and ease of use are affected by error type and correction method:

- Baber, 1991
- Bekker et al., 1995
- Brems et al., 1995
- Cohen and Mandel, 1982
- Fink, 1981
- Jones et al., 1992
- Kamm, 1994
- Martin, 1976
- Peckham, 1984
- Peckham, 1986
- Pelton, 1993
- Quinnell, 1995
- Schafer, 1994
- Schurick, Williges, and Maynard, 1985
- Simpson et al., 1985, 1987
- Spine, Maynard, and Williges, 1983
- Wattenbarger et al., 1993

Experiments performed by Schurick et al. (1985) and Spine et al. (1983) manipulated error correction methods (command vocabulary) to evaluate system performance and user acceptance. In general, their results demonstrated that user acceptance and task performance were influenced by error-correction command vocabulary and the type of feedback provided. Brems et al. (1995) employed natural language reprompting strategies for correcting rejection errors made by an ASR system, and found improvements in task performance time using natural language feedback messages versus traditional error messages.

Feedback. Some form of feedback is necessary to detect and correct recognition errors, or to inform the user that a command has been executed by the ASR system. Research has shown that the appropriate level and modality of feedback are dependent on the tasks being performed by the user of a speech-input system and can significantly affect system performance (Baber et al., 1992; Forren and Mitchell, 1986; Schurick et al., 1985; Steiner et al., 1983). If the user is performing a primary visual-task which is augmented by speech recognition, auditory feedback may interfere less with primary task performance and be more acceptable to the user than visual feedback. If error correction is difficult or execution of the task is safety-critical, secondary feedback is expected to improve task performance by minimizing errors. The following references discuss feedback provided by ASR systems and how task performance and ease of use are affected by feedback level and feedback modality:

- Baber, 1991
- Baber et al., 1992
- Brems et al., 1995
- Cohen and Mandel, 1982
- Costet, 1982
- Forren and Mitchell, 1986
- Jones et al., 1989, 1992
- Kamm, 1994
- Martin, 1976
- Peckham, 1986
- Schurick et al., 1983, 1985
- Simpson et al., 1985, 1987
- Steiner et al., 1983
- Wattenbarger et al., 1993

Speaking mode of user. The speaking mode required by an ASR system can significantly affect user acceptance and task performance. Task completion time, for example, increases when users have to pause between each word in a string of commands while using an isolated-word recognizer. In addition, it is unnatural for users to pause briefly between each spoken input or speak with consistent intonation as is required by isolated- or connected-speech recognition systems. However, a continuous-speech recognizer allows the user to speak in his or her natural speech rhythm and intonation without having to pause between each spoken input. Task completion time decreases compared to an isolated-word recognizer. The problem is that current continuous-speech recognition systems are unable to demonstrate the same high levels of recognition accuracy as isolated-word speech recognition systems. The following references discuss ASR system speaking mode, how speaking mode affects other system parameters, and how task performance and ease of use are affected by the speaking mode required:

- Bennett et al., 1989
- Biermann, Fineman, and Heidlage, 1992
- Fink, 1981
- Furui, 1994
- Gould, Conti, and Hovanyecz, 1983
- Kamm, 1994
- Lea, 1980b
- Martin, 1976
- Martin and Welch, 1980

- Quinnell, 1995
- Schafer, 1994

Experiments performed by Biermann et al. (1992), and Gould et al. (1983) manipulated speaking mode to evaluate system performance and user acceptance. In general, the results demonstrated that speaking mode was marginally significant in affecting task completion times, with continuous-speech being faster than isolated-speech for a listening typewriter task (Gould et al., 1983), and user acceptance decreased when switching from an isolated-word to a connected-word natural language editor (Biermann et al., 1992).

Vocabulary size. The size of the active vocabulary can affect ASR system performance and user acceptance. In general, recognition accuracy and recognition speed decrease while recognition errors increase as the size of the active vocabulary is increased. However, with the use of appropriate grammars and careful selection of command vocabulary, the size of the active vocabulary at any given point in the interaction can be maintained at levels easily handled by the speech recognition system. Research has been performed to determine the vocabulary size required by specific applications such as graphic manipulation, automated call-routing, and inventory recording (Hauptman and McAvinney, 1993; Wharton, Marics, and Engelbeck, 1993; Zoltan-Ford, 1984). Subjects were allowed to use any words that might accomplish the task. The results indicated that a large number of vocabulary words are generated when subjects are allowed to use any word to complete a task. However, this large vocabulary can be reduced by selecting the common root-phrases of vocabulary words used to perform specific tasks (Wharton et al., 1993) and by forcing users to model the vocabulary and phrase structure of the feedback provided by the system (Zoltan-Ford, 1984). The vocabulary size should be sufficient to allow the user to say all common words related to the task being performed, but should not be too large that the user is unable to remember all the vocabulary commands required to successfully complete the task. The following references discuss ASR system vocabulary size, how vocabulary size affects other system parameters, and how task performance and ease of use are affected by the size of the active vocabulary:

- Atal, 1994
- Baber, 1991
- Bennett et al., 1989
- Bernold, 1993
- Casali et al., 1990
- Cohen and Oviatt, 1994
- Costet, 1982
- Fink, 1981
- Furui, 1994
- Gould et al., 1983
- Lea, 1980a
- Lee et al., 1990
- Markowitz, 1995
- Pallett, 1986
- Pelton, 1993
- Poock, 1986
- Schafer, 1994
- Spine, Williges, and Maynard, 1984

Experiments performed by Casali et al. (1990), Gould et al. (1983), and Spine et al. (1984) manipulated vocabulary size to evaluate system performance and user acceptance. In general, the results demonstrated that vocabulary size significantly affected task performance. Task completion times were faster when larger vocabularies were available (Casali et al., 1990; Gould et al., 1983). However, available vocabulary did not have a major effect on user acceptance of the ASR system (Casali et al., 1990).

Visual, Manual, and Cognitive Demands of In-vehicle Tasks

The driver of an automobile performs a series of complex tasks. The primary driving task consists of gathering visual information from the forward scene, processing that information, and performing appropriate manual responses to maintain the vehicle on the road and avoid collisions with other objects. The driver may also perform secondary tasks while concurrently driving the vehicle. These secondary tasks may include obtaining information about vehicle performance and status, responding to changes in driving conditions, or accommodating comfort and entertainment interests of vehicle occupants. In general, secondary tasks place similar demands on the driver as the primary driving task; most secondary tasks require visual, manual, and cognitive resources to be allocated so the task can be successfully completed.

Numerous studies have evaluated various aspects of the driving task with regard to attentional demand. Attentional demand when driving a vehicle can be divided into three categories: visual demand, manual demand, and cognitive demand. Driver attentional demand has been operationally defined by driver eye-movement behavior, hands-off-wheel time, and other driver-vehicle performance measures (Dingus, Antin, Hulse, and Wierwille, 1989). Each category of driver attentional demand is analogous to a micro-measurement of driver workload. The visual, manual, and cognitive demands of current in-vehicle secondary tasks are discussed in this section and, whenever possible, comments are made regarding driving performance and user acceptance of secondary-task design as related to attentional demand.

Visual demand. Vision is the primary sensory modality for gathering information from the driving environment. Inside most vehicles, the auditory channel is used only to alert the driver to warning information (e.g., door ajar, seat-belt not fastened, headlights left on, etc.). However, the majority of in-vehicle stimulus information is presented to the driver through the visual channel.

Visual demand refers to the extent to which foveal vision is required to obtain information from the forward road scene or an in-vehicle display or control. The driver of a vehicle has only one foveal-vision resource from which he or she can obtain detailed information about a single location in the driving environment at any given time (Wickens, 1992; Wierwille, 1993). Therefore, the driver must time-share visual resources between the forward road scene and other in-vehicle tasks that require vision (Wierwille, 1993). Visual demand imposed on the driver by a secondary task can affect driving performance, behavior, and acceptance of secondary-task design.

Manual demand. The primary response modality used in the driving environment is through the manual channel. The driver performs manual outputs to control the vehicle and accomplish in-vehicle tasks. The hands are used to steer the vehicle and control in-vehicle systems such as the radio or the climate control system, while the feet are used to control the acceleration and deceleration of the vehicle. Unlike the visual channel, two or more different tasks can be performed simultaneously through the manual channel (e.g., two hands and two feet can be used to both control the vehicle and perform in-vehicle tasks). A driver can keep one hand on the steering wheel to control the vehicle while the other hand activates the windshield wipers. However, occasions may arise when manual output must be time-shared (Wierwille, 1993). For example, both hands are usually required to manipulate the steering wheel when making a sharp turn. Manual control of a secondary task would be interrupted to maintain good control of the vehicle when the driver is forced to make a sharp turn. Wierwille (1993) states that drivers must time-share visual inputs continuously and only occasionally time-share manual outputs.

Manual demand refers to the extent to which manual response is required to perform an in-vehicle task. Steering the vehicle has a high level of manual demand; however, the primary task of driving is to maintain directional control. Depressing the horn button or setting the directional signal lever is a secondary task which has a low level of manual demand. Manually dialing a handheld cellular telephone is a secondary task which has a high level of manual demand and may interfere with the primary task of safely steering the vehicle. The manual demand imposed on the driver by a secondary task can affect driving performance, behavior, and acceptance of secondary-task design.

Cognitive demand. Eye-movement behavior is sometimes used to explain the cognitive demands of driving (e.g., the internal expectancies and understandings that drive attention). However, the cognitive demand of a secondary task is more difficult to ascertain than a task's visual and/or manual demand. The difference between cognitive demand and visual demand is best explained through an example.

Eye-movement behavior is recorded along with other measures of driving performance to evaluate the effects of two different types of automotive head-up displays (HUDs) on driver behavior and performance. The first HUD has low information content; for example, speed information combined with text-only turn-by-turn navigation information. The second HUD has high information content; for example, speed information combined with full-color moving-map navigation information. Analysis of the eye-movement data shows no significant difference between the two HUDs. Glance duration and glance frequency do not statistically differ between the two display types. At first, a researcher might conclude that the two displays have the same cognitive demand. However, upon analysis of the other measures of driving performance, a significant difference in response time to a forward scene event exists between the two displays. When driving the vehicle equipped with the high-content navigation HUD, driver response times to the forward scene event were significantly longer than response times when driving the low-content navigation HUD vehicle. Visual demand (as measured by glance duration and glance frequency) did not differ between the two displays, but cognitive demand (as measured by response time) differed significantly. This difference may be explained by the concept of cognitive capture. When a driver fixates the low-content HUD and mentally processes its information, enough spare cognitive resources may exist that the driver is able to attend to information presented in the forward scene. However, when a driver fixates the high-content HUD and mentally processes its information, spare cognitive resources are non-existent. Therefore, the driver is unable to attend to informational events presented in the forward scene.

This example demonstrates that the cognitive demand imposed on the driver by a secondary task can affect driving performance, behavior, and acceptance of secondary-task design. Since manual demands are generally not time-shared and cognitive demands are difficult to measure, Wierwille (1993) believes the visual demand of in-vehicle tasks to be the most important consideration for the design and evaluation of automotive secondary tasks.

Experimental data. This section describes empirical research that evaluated the visual, manual, and/or cognitive demands of in-vehicle secondary tasks performed by the driver. The type of demand evaluated, the secondary tasks performed, and the experimental results are briefly explained. Measures of driving performance and user acceptance are reported when available. The references are presented in chronological order.

The effect of performing in-vehicle secondary tasks on visual, manual, and/or cognitive demands has been researched as early as 1969. Brown, Tickner, and Simmonds

(1969) evaluated the concurrent use of a radiophone (early version of the cell phone) while driving a vehicle on a closed course. They claimed two main sources of interference between driving and using a phone: (1) having to hold the headset and manipulate push buttons to send or receive calls (manual demand); and (2) having to switch attentional demand between the visual and auditory channels while talking on the phone (cognitive demand). Subjects drove an experimental vehicle around a test track while concurrently performing a telephoning task. Judgments of "gap size" (which participants had to decide whether to drive through a gap between barriers or around the barriers) were recorded to evaluate the cognitive demand of the telephoning task. The results indicated that when subjects operated under conditions of divided attention (telephoning while driving), they attempted to drive through far more gaps that were smaller than the car and slightly fewer that were larger. Brown et al. concluded that performing a telephone communication task while driving significantly affects cognitive demand.

Kames (1979) evaluated the effects of various telephone keypad designs and mounting locations on driving performance. Visual, manual, and cognitive demands were evaluated for the various telephone control unit designs and were compared with the attentional demands of adjusting the radio. Subjects performed a series of pre-trained telephone dialing tasks using each combination of telephone dials (rotary or push-button), push-button arrays (4 x 3 or 6 x 2), and mounting locations (handheld, on the dash, and near the visor). Throughout the drive, subjects were instructed to respond to the appearance of a light in the forward field of view. Response time to the illumination of the light provided a measure of situational awareness, or workload, imposed by the telephone dialing task. Subjects were also instructed to make adjustments to the radio inside the vehicle. The results indicated that the design of a telephone dial that is within the reach and sight of the driver has little effect on the visual, manual, and cognitive demands of driving. User preferences favored the 4 x 3 push-button array mounted on the dash. Interestingly, no significant differences in driving performance measures were found between performing telephone dialing tasks and adjusting the radio.

Mourant, Herman, and Moussa-Hamouda (1980) evaluated the effect of in-vehicle control location on visual and manual demand. Drivers performed a series of in-vehicle tasks using both stalk-mounted and panel-mounted controls. Measures of glance duration and glance frequency were recorded to evaluate visual demand, and measures of movement time (the time from when the subject's hand left the steering wheel until a control was touched) and operation time (the time from when the subject's hand touched the control until the control was activated) were recorded to evaluate manual demand. Control type and location were found to significantly affect visual and manual demand. As the distance to the control increased, visual demand and manual demand increased (e.g., glance frequency and movement time increased). No mention of the effect of control location on tracking task performance (which was concurrently performed by subjects when interacting with the various controls) was reported.

Haller, Bouis, and Heintz (1981) evaluated the effect of trip computer keyboard design on visual and manual demand. Drivers had to enter 3- or 4-digit numbers or select a trip computer function while concurrently driving a simulator. The keyboard designs used were a 4 x 3 telephone keypad and a sequence keypad (e.g., each key corresponds to a digit place and digits are entered by successive presses of the corresponding key place). Measures of eye-movement behavior, task-completion time, and tracking error were recorded to evaluate visual and manual demand. Keypad design significantly affected visual and manual demand. Task completion times were significantly longer for the sequence keypad, indicating greater manual demand. However, eye fixation times and tracking error were significantly less for the sequence keypad design. The telephone keypad had greater visual demand, which affected driving performance compared with the sequence keypad.

Zwahlen and DeBald (1986) evaluated the effect of visual demand on driving performance. Subjects drove a test vehicle down the center of a closed airport runway while performing one of three tasks: (1) driving with eyes fixating the road ahead, (2) driving with eyes closed, and (3) driving while reading text inside the car. Lateral deviation from the runway's centerline was recorded at equal intervals for each experimental condition. The results showed that deviation from the runway centerline increased at an exponential rate as the distance from the task starting point became greater, when eyes were fixated inside the vehicle. Zwahlen and DeBald concluded that increasing the visual demand of in-vehicle tasks--in particular, the single glance durations required to perform the task--would result in increased lane deviations exhibited by drivers, which would increase the probability of an accident.

The visual, manual, and cognitive demands of an in-vehicle navigation system were evaluated by Dingus et al. (1989). Drivers performed a series of navigational tasks using an in-vehicle computerized navigation system, and they also performed several conventional automotive tasks. Conventional tasks were included so comparisons could be made between navigational and conventional tasks in terms of attentional demand required. Three categories of measures were used to determine attentional demand: eye-movement measures (visual demand), task completion measures (manual demand), and driving performance/behavior measures (cognitive demand). Several tasks using the navigation system were found to significantly increase visual demand when compared with conventional tasks (e.g., tuning the radio). However, cognitive demand (as measured by driving behavior and performance) was not significantly affected by task performance. The results of task performance on manual demand were not presented in this paper; however, it was reported that manual demand of most navigation tasks was comparable to that of one or more conventional tasks (e.g., adjusting audio system, adjusting climate control, etc.).

Kurokawa (1990) performed a series of experiments to evaluate the effects of instrument panel (IP) control and display type, instrument panel macroclutter/microclutter, and instrument panel control labels on driving performance. Subjects performed a series of in-vehicle display and control tasks while driving a simulator. Measures of eye-movement behavior and task completion time were recorded to evaluate visual and manual demands. In general, visual and manual demand were significantly affected by control-display type, IP macroclutter/microclutter, and control labeling. The results from the experiments were used to develop a computer program to quantitatively evaluate current and proposed instrument panel designs.

Brookhuis, de Vries, and de Waard (1991) evaluated the effects of cellular phone use on driving performance. Measures of lateral position and steering wheel movements were recorded to evaluate manual demands, and measures of "rearview mirror glances" were recorded to evaluate visual demands. Subjects performed a series of conversational tasks and dialing tasks under various traffic conditions using both a handheld and hands-free telephone. The results demonstrated that carrying on a telephone conversation while driving affects driving performance. In terms of lateral position, talking on the telephone while driving significantly decreased lane deviations. In other words, drivers tended to swerve less when talking on the phone compared to driving without using the phone. However, dialing the telephone had a negative effect on driving performance. Having to dial the telephone by hand while driving showed a significant effect on steering wheel movement when compared to answering a call manually. Brookhuis et al. concluded that using a cellular telephone while driving (actually carrying on a conversation) had a nominal effect on driving performance. An exception to that conclusion is that driving performance was severely degraded when the driver was manually dialing the telephone. The researchers recommend that only hands-free cellular telephones, preferably with voice-activated dialing, be allowed for use in automobiles. Visual demand as measured by rearview mirror glance frequency was not significantly affected by the telephone conversation task.

Fairclough, Ashby, and Parkes (1993) evaluated the visual demand of two in-vehicle navigation devices. Measures of glance duration and glance frequency were used to calculate total glance durations to the navigation devices while driving. Their results indicated that driver visual demand was sensitive to the type of route-guidance information presented (map vs. text); map displays demonstrated larger total glance durations than text displays. In addition, the results demonstrated that drivers allocated more visual resources to the navigation displays at the expense of other areas in the driving scene.

Research performed at the University of Michigan Transportation Research Institute (UMTRI) investigated car phone design and usability (Serafin, Wen, Paelke, and Green, 1993). Two types of phones (manual and voice-operated) and two types of displays (instrument panel and head-up display) were experimentally tested using a driving simulator. Measures of lane position and dialing time were recorded to evaluate manual demand. Subjective data was also collected to evaluate which design was preferred. Subjects were instructed to perform a series of dialing tasks which included dialing familiar and unfamiliar 7- and 11-digit numbers while concurrently driving the simulator. All combinations of phone type, display type, type of phone number, and length of phone number were evaluated by each subject. Significant main effects were found for input method, task, and age. Driving performance was most affected during the period of dialing the phone. Voice input resulted in significantly smaller deviations in lane position than manual dialing. Younger drivers performed significantly better than older drivers during the dialing and driving condition. Dialing time was also significantly affected by all variables except display type. Upon evaluation of the interactions, it was determined that voice-input resulted in significantly faster dialing times for older subjects than the manual phone, and that all subjects dialed unfamiliar numbers significantly faster with the voice-input phone than the manual phone. Serafin et al. concluded that use of voice-input cellular phones would improve traffic safety compared with conventional manual-input phones. Further, subjective driver preference ratings indicated that the voice-HUD combination was the best.

Dingus (1995) analyzed eye-movement behavior from two studies of in-vehicle navigation displays. He compared the visual sampling behavior (visual demand) while using a navigation system with the visual sampling behavior (visual demand) while driving without a navigation system. From the analysis, Dingus determined that using an in-vehicle navigation system drew visual attentional resources away from both the forward and left/right roadway scanning resources.

Kiger, Rockwell, and Tijerina (1995) performed an on-the-road study to determine heavy-vehicle driver visual demand under normal operating conditions. Professional truck drivers drove a heavy vehicle over a series of roadway types and lighting conditions while performing in-vehicle tasks (e.g., check gauges, tune radio, etc.). Eye-movement behavior was recorded and used in the development of baseline measures of visual allocation during normal operating conditions. Road type and in-vehicle tasks were found to be significant factors affecting visual demand.

Tijerina, Kiger, Rockwell, and Tornow (1995) performed an on-the-road study to assess the workload imposed on heavy-vehicle drivers by an in-cab text messaging system and a cellular phone. Professional truck drivers performed a series of text messaging and cellular phone tasks while driving. Measures of visual demand (eye-movement behavior), manual demand (driving behavior and performance), and cognitive demand (driving behavior and performance) were recorded to evaluate driver workload while performing in-vehicle tasks. Their results demonstrated that the type of in-vehicle task performed significantly affected visual, manual, and cognitive demands. Interestingly, manual cellular phone dialing required the same visual, manual, and cognitive demands as manually tuning the radio.

This review of the literature demonstrates that in-vehicle task performance can affect the attentional demand of the driver. It is also important to understand that visual, manual, and cognitive demands can be affected by the driving environment. Wierwille and his colleagues have performed a series of experiments that evaluated the effects of driving task demand on performance of secondary in-vehicle tasks. In general, the results indicated that as driving task demand increases (e.g., road curvature increases, traffic density increases, etc.), drivers exhibit changes in their strategies for handling the visual, manual, and cognitive demands imposed by in-vehicle displays and controls. A review of this research and the results can be found in Wierwille (1993).

Implications for Incorporating ASR Technology in Automotive Control Systems

The review of the literature in the last section provided insight into the actual visual, manual, and cognitive demands placed on the driver by current in-vehicle interface designs. Note that the majority of in-vehicle tasks described were performed manually by the driver. In the few instances where speech was used as an input method, the primary driving-task performance was not affected or improved when compared with similar manual input tasks. This section describes the type of ASR system that should be used to control in-vehicle secondary tasks and discusses what improvements in driver performance and user acceptance may be expected. This section also describes how in-vehicle secondary task functions can be selected for control using speech input.

ASR technology and in-vehicle secondary tasks. The use of speech recognition technology as an additional input channel to control secondary-task functions may produce positive effects on driving performance, behavior, and acceptance of secondary-task design. As was demonstrated in the literature on in-vehicle task demands and the discussion on attentional resource theory and S-C-R compatibility, manual control of secondary tasks can adversely affect performance of the primary driving task. This section attempts to provide recommendations about what levels of ASR system performance are required before any secondary task is considered for control through speech input. Some speech recognition parameters will require empirical evaluation before appropriate levels of system performance can be specified. It is important to note that recommendations about ASR system performance are based on human factors principles. Cost or technological limitations are not a consideration in selecting any system variables.

Any speech recognition system being considered for use in an automobile should have the following parameters:

(1) The system should be speaker adaptive, or at least speaker independent. Vehicle owners will probably not want to take the time to enroll a speaker-dependent system. It does not take a new-car owner long to figure out how to manually control all the secondary systems in the vehicle, so why should he or she be required to spend an hour or more training an ASR system to control just the audio system? In addition, an automotive ASR system may have many different users, all of whom would require enrollment if a speaker-dependent system is to work reliably. The key issue for the implementation of speech recognition technology inside the automobile is to make the interaction with the system as transparent to the user as possible. One way this is accomplished is through the selection of a speaker-adaptive system. Both speaker-independent and speaker-adaptive ASR systems allow multiple users to speak commands to the system without any prior enrollment. The speech variability between potential users is best accounted for by

speaker-adaptive systems because these systems are constantly adjusting and updating algorithms to match variations in speech patterns. A speaker-adaptive system should also perform more reliably under conditions where drivers experience stress or fatigue, which can increase intra-speaker speech variability.

(2) The system should allow continuous speech input from the user and incorporate a key-word-spotting grammar. The driver of an automobile should not be forced to pause between each command or speak in a monotone fashion as would be required by isolated-word and connected-word speech recognizers. To make the interaction with an ASR system transparent, the user should be allowed to speak in a natural, conversational dialogue. Continuous-speech recognizers allow users to do that. The use of a key-word-spotting grammar provides freedom to word command phrases in a way that seems natural to the user. The only restriction is that the command phrase must contain recognizable vocabulary items to successfully complete the task. A speech-understanding system would be the ultimate choice for use inside the automobile. However, until the technology matures to the level where a speech-understanding system would perform reliably in the automotive environment, a continuous speech input ASR system with key-word spotting is the next best choice.

(3) The size of the command vocabulary should be large enough to allow users to say all common words related to the task being performed, but not too large that users are unable to remember all the vocabulary commands required to successfully complete the task. The use of a finite-state grammar and careful selection of the command vocabulary can reduce the size of the active vocabulary at any given point in the interaction to a level that can be easily handled by the speech recognition system and remembered by the user.

The following speech recognition parameters require further empirical testing before appropriate levels of system performance could be specified for use in an automobile. First, an appropriate level of recognition accuracy must be empirically determined so that driving performance, ease of use, and user acceptance are maintained at, or exceed, levels comparable to current manual interaction methods. It was hypothesized that an automotive ASR system need not perform with 100% recognition accuracy. A lower recognition accuracy may be sufficient to perform in-vehicle tasks while maintaining acceptable levels of driving performance, system usability, and user acceptance. The empirical research incorporated a range of recognition accuracy levels that were believed to significantly influence driving performance, ease of use, and user acceptance. By including a broad range of accuracy levels, guidelines and recommendations can be made that allow designers to determine adequate levels of ASR system performance (e.g., recognition accuracy) before considering such an input technology.

Second, the types of recognition errors common to an ASR system must be empirically evaluated to determine the effect that different error types may have on driving performance, ease of use, and user acceptance. The types of recognition errors common to a system are highly dependent on the recognition accuracy of that system. The ultimate goal is to minimize substitution and insertion errors. Substitution and insertion errors typically require additional steps in the interaction to correct the error. Rejection errors, however, are more easily tolerated by users--particularly if they occur infrequently. It should be intuitive that one way to reduce recognition errors (in particular, substitution and insertion errors) is to increase the recognition accuracy and discrimination accuracy of the ASR system. However, tradeoffs may exist that allow lower levels of recognition accuracy and discrimination accuracy but still maintain acceptable levels of driving and user performance. Recognition error type was evaluated in this research by allowing different error types to occur during interactions with the ASR system and measure the effects of each error type on driving performance, ease of use, and user acceptance. The results allow designers to select command vocabulary and ASR equipment (e.g., microphones, recognition software, etc.) that minimize error types that were found to negatively affect the performance measures of interest.

Third, feedback level and modality provided by the ASR system must be empirically determined so that an appropriate level and modality is selected for use in automobiles. The level (primary or secondary) and modality (auditory and/or visual) of feedback should optimize driving performance, ease of use, and user acceptance when a driver interacts with an ASR system. It was hypothesized that an ASR system that provides auditory-only or auditory-and-visual feedback should produce the highest levels of driving performance, ease of use, and user acceptance. However, it was believed that feedback level and modality might be task-function dependent. Primary feedback may be sufficient for certain task functions, while secondary feedback may be required by others. Feedback level and modality were evaluated in the research by testing the effect of different feedback combinations on measures of driving performance, ease of use, and user acceptance. The results allow designers to select appropriate feedback levels and modalities for various task functions being considered for control through speech input.

Finally, the method of error correction allowed by the ASR system must be empirically tested to determine what error correction method optimizes driving performance, ease of use, and user acceptance. Error correction methods may depend greatly on the recognition accuracy of an ASR system, the types of recognition errors common to the system, and the type of feedback provided by the system. Therefore, it was important to know how recognition accuracy, error types, and feedback interact with the method of error correction to affect driving performance, ease of use, and user acceptance to determine an optimal error correction method.

Now that the important ASR system parameters have been specified, appropriate task functions were selected for control via speech input. Vail (1986) generated four categories of driving functions which were used in evaluating the utility of voice-activated controls for physically handicapped drivers. The four categories were defined as follows.

(1) *Critical driving functions.* Functions that are critical to the operation of the vehicle and completion of the driving tasks, and can occur while the vehicle is or is not in motion (e.g., gear shift, steering, speed control, etc.).
(2) *Normal driving functions.* Functions that aid in the operation of the vehicle, can increase driver comfort/performance, and can occur while the vehicle is or is not in motion (e.g., windshield wiper, window, climate control system, navigation system, etc.).
(3) *Luxury driving functions.* Functions that are not critical to the operation of the vehicle or completion of the driving task, and mainly exist for the driver's enjoyment (e.g., radio, cell phone, etc.).
(4) *Single performance functions.* Functions that are performed only once during driving and while the vehicle is not in motion (e.g., ignition switch, trunk release, gas door release, etc.).

These four categories provide information about the major function allocation issues that need to be considered for implementing ASR technology into an automobile. They also provide a starting point from which analysis of secondary-task function allocation can begin. Functions that are critical to the driving task should not be considered for control via speech input because ASR technology will probably never achieve the capability to execute such important task functions as are currently performed through manual input by the driver. However, the following categories of secondary-task functions are worth considering to be controlled through speech input: (1) normal driving functions, (2) luxury driving functions, and (3) single operation functions.

Method for selecting ASR-controlled secondary-task functions. Function or task allocation is defined as the division of individual tasks between members of an organization and between people and machines (Kirwan and Ainsworth, 1992). The process of function allocation for an automobile consists of determining which in-vehicle functions should be performed by the human operator and which functions should be performed by the automobile. This taxonomy can be taken a step further by determining which in-vehicle functions should be manually controlled by the driver, and which functions should be controlled through speech input from the driver. Once function allocation is completed, the extent of operator involvement in the control of a system is defined (Kirwan and Ainsworth, 1992). The division of functions between person and machine (in particular, the driver of an automobile and an ASR system) is a primary focus of this research.

The systems engineering approach to the design and development of complex systems formally assigns the role of function allocation to the "concept" phase of the design lifecycle. However, function allocation takes place throughout the entire design process and is iterative (Bailey, 1989). The assignment of functions to either the machine or the human operator is usually determined by the superiority of one over the other in performing the task or because economic considerations dictate a particular assignment. However, situations arise where a function may be performed equally as well by the machine as it is by the human operator. The designer is now given the task to decide how this function will be accomplished.

Due to constraints imposed by cost, technology, and design characteristics, designers are unable to create system interfaces that are optimal in every way. Tradeoffs are made between design criteria to meet specific design objectives. It is the designer's job to determine what combination of criteria best satisfies the design objectives. Fortunately, designers have decision tools to help them identify the strengths and weaknesses of proposed design solutions and to serve as structured methodologies for maximizing the design advantages. A decision tree was generated as part of this research effort to determine which in-vehicle task functions are amenable to ASR technology. The decision tree helped determine the task functions evaluated in the empirical research and provided input for the final project objective of usable design guidelines.

A subset of in-vehicle task functions had been identified that were determined by the decision tool to be either amenable or not to speech recognition technology. This subset of task functions was combined with the set of critical speech recognition variables to determine how usability, driver performance, and driver preference/acceptance are affected by in-vehicle tasks using speech recognition for control of secondary-task functions.

DECISION TREE ANALYSIS

Decision trees are one type of tool that can be used to facilitate the decision making process (Kleinmuntz, 1987). The decision problem is represented in the form of a tree with numerous branches and endpoints (sometimes referred to as twigs). The decision tree contains points of decisions where designers or engineers are required to decide the most likely outcome for a given situation. The tree also contains points of action which represent the recommended course of action given a series of decisions (e.g., to allocate a given in-vehicle task function to be controlled using speech input). The decision points and action points are connected one by one to form branches and endpoints that represent various phases of the decision making process, leading to the final result of task-function allocation. Decision criteria such as safety, usability, and driver preference/acceptance provide a logical basis for determining whether a given task-function should be considered for control using speech input or some other conventional input method.

A series of decision trees were developed in this research effort to provide designers and engineers of in-vehicle tasks with an analytical methodology for determining whether a given task should be performed using manual controls or using speech input (see Appendix A). The manner in which this decision tool is to be used is as follows.

First, a task function is selected from the system under consideration. For example, selecting a radio station from a series of preset stations (or stations stored in memory) is a task function of the audio system. The task function becomes the foundation or reference used to answer questions found at each node of the decision tree. When a designer or engineer is asked a question by the decision tree, the answer should be based on the characteristics of the task function and not on characteristics of the overall system. The system and task function are entered into a function allocation chart (see Table 2), which is used in justifying the overall final decision of whether the specific task function should be performed using manual input or automatic speech recognition (ASR) input.

Table 2. Example of a task-function allocation decision chart.

System	Task Function	Driving-Function Category	Behavioral Category	Behavioral Allocation Decision	Safety Allocation Decision	Usability Allocation Decision	Driver Preference/ Acceptance Allocation Decision
Audio	Select preset	Luxury	Selective Activation	ASR Control	ASR Control	Manual Control	Both ASR and Manual Control
Audio	Adjust volume	Luxury	Simple Locate and Reach/Fine Adjustment	Manual Control	Manual Control	Manual Control	Manual Control
Climate	Adjust fan speed	Normal	Simple Locate and Reach/ Gross Adjustment	ASR Control	Manual Control	Manual Control	Both ASR and Manual Control
Cell Phone	Dial phone number	Luxury	Selective Activation	ASR Control	ASR Control	ASR Control	ASR Control

Second, the task function of interest is defined according to its driving function category (described above and in Vail, 1986). This is accomplished through answering the questions found in Figure A-1, the "Driving Functions" decision tree (see Appendix A). Defining the task function according to a specific "driving function" helps guide the user of the decision tool in making appropriate decisions to questions asked later in the decision tree process. For example, the answer to the question "Is an immediate and correct response required when performing the task function?" is most likely different for task functions defined as *normal* driving functions compared with task functions defined as *luxury* driving functions. Once the task function is defined according to a driving function category, it is entered into the function allocation chart (see Table 2).

Next, the behavioral characteristic(s) of the task function is determined by answering the question found in Figure A-2, the "Behavioral Categorization and Decision Allocation" decision tree (see Appendix A). Again, by further defining the task function by its behavioral characteristic(s), the user of the decision tool is grounded in making better decisions to questions asked later in the decision tree process. The behavioral category of the task function is also entered into the function allocation chart (see Table 2).

The behavioral categories were adapted from Kurokawa (1990). Kurokawa developed a software program that evaluates automotive instrument panel design. Using the program, a designer or engineer would enter information about a given instrument panel design, such as the instruments used and their interface characteristics (e.g., type of controls and displays contained in each instrument), or the person could enter the behavioral characteristics associated with a given instrument panel design. The software would determine a "figure of merit" for the given instrument panel design which could be compared with alternative designs. The point of interest from this research is the categorization of in-vehicle tasks into behavioral elements. Categorizing in-vehicle tasks using behavioral characteristics provides an instrument-independent method for describing in-vehicle tasks. Explanations and examples of each behavioral category are as follows:

(1) *Simple Locate and Reach.* Drivers locate a target object, reach for it, and manipulate it (e.g., driving lights control).
(2) *Manual Delay.* This includes tasks that require a control to be held for a period of time (e.g., power window control).
(3) *Selective Activation (1 of N Controls).* Drivers locate and reach for a target control and make a selection from a group of controls that the instrument offers (e.g., radio station preset controls).
(4) *Gross Adjustment.* The adjustment process, after locating and reaching a target control, involves feedback, which requires greater attentional demand from the driver than in category 3, but less than in category 5 (e.g., fan speed control).
(5) *Fine Adjustment.* This adjustment process requires iterations of adjustment input and feedback to achieve a desired state from many discrete states or from a continuum (e.g., temperature control or volume control).

After defining the task function according to its behavioral category, a decision allocation is made by answering the questions found in Figure A-2, the "Behavioral Categorization and Decision Allocation" tree (see Appendix A). The allocation decision from Figure A-1 or Figure A-2 is entered into the function allocation chart under the heading of "Behavioral Allocation Decision" (see Table 2).

The final step is for the designer or engineer to use the remaining decision trees, each of which independently addresses the issues of safety, usability, and driver preference/acceptance, to make allocation decisions regarding the task function of interest.

These decision trees are shown in Figures A-3, A-4, and A-5, respectively (see Appendix A). The allocation decisions from each tree are entered under the appropriate heading in the function allocation chart (see Table 2).

The task-function allocation decision chart (see Table 2) summarizes the function allocation decisions for each task-function as determined by the various decision trees. The chart is useful because it allows the designer or engineer to view task-function allocation decisions based on orthogonal criteria of interest (i.e., safety, usability, and driver preference/acceptance). The chart can be used in making the final "input method" allocation decision for a given task function. Using Table 2 as an example, it is obvious that the task function "adjust volume" for the audio system should be performed using manual input; all allocation decisions suggest manual control. However, the task function "adjust fan speed" for the climate system has two allocation decisions suggesting ASR control, one allocation decision suggesting manual control, and one suggesting both ASR and manual control. The designer or engineer may want to consider having fan speed controlled using ASR technology only if safety consequences outweigh task function usability and driver preference/acceptance of task function design. If other task functions within the climate system are recommended for ASR control, then it may be desirable to control fan speed using ASR control only or a combination of ASR and manual control.

The decision trees and task-function allocation decision chart are not the final word in the decision process of whether a task function should be performed manually or using speech input. Other factors, such as cost and manufacturability, must also be considered in the final decision. The decision trees and chart are tools that provide task-function input-modality allocation guidelines based on human factors principles and consideration for the end-user of the product (i.e., the automobile customer). The decision tool is used for determining whether an in-vehicle task function should be performed using speech input or manual input. The results from the empirical research are used for determining the appropriate ASR system parameters (i.e., system feedback and error correction method) and the appropriate levels of ASR system performance (i.e., recognition accuracy and recognition errors allowed) necessary before implementing ASR technology inside the automobile to control an in-vehicle task. The decision tool combined with the results from the empirical research should help improve in-vehicle task design and reduce the burden on designers and engineers to implement technology simply because it exists, instead of implementing technology to improve end-user performance and end-user satisfaction.

SUMMARY

Overview

The literature review provides insight into human information processing and how it may be affected by system design. Interactions with complex systems (in particular, dual-task performance) can be improved if the way in which information is presented is carefully matched to an appropriate response channel for the given task. For example, the literature on attentional resource theory has shown that tasks that are spatial in nature, such as driving an automobile, are performed best with manual type responses or inputs, while tasks that are verbal in nature, such as entering navigation waypoints or radio frequencies, are performed best with verbal (vocal)-type responses or inputs. This is especially true when the two tasks are performed concurrently. If the driving task and a tune-radio task are performed concurrently using manual inputs, dual-task performance is expected to be worse than if the tune radio task was performed using verbal input. Matching the information presentation (or stimulus) modality, and the central processing modality of the information presented with the appropriate response modality should increase stimulus-central processing-response (S-C-R) compatibility and decrease mental workload for the operator of the system. Increasing S-C-R compatibility and decreasing mental workload generally results in improved operator performance and user satisfaction with the system.

The literature review also describes the important parameters associated with automatic speech recognition (ASR) technology and how it may affect user performance and satisfaction of automobile systems incorporating such technology. This review includes (1) a description of speech recognition parameters and their definitions; (2) the barriers--both technological and behavioral--to implementing ASR technology inside the automobile (e.g., recognition accuracy in noisy environments; type and modality of system feedback; error correction method allowed; size of the command vocabulary; system's capability to recognize speaker-independent input; etc.); (3) the important ASR system parameters to be evaluated through empirical research; (4) a review of the attentional demands of in-vehicle tasks; and (5) the implications and potential benefits of incorporating ASR technology to control in-vehicle tasks. The literature review on ASR technology provides initial guidelines as to what should and should not be included in a speech recognition system for controlling in-vehicle tasks. The review also allowed the number of variables associated with speech recognition systems to be narrowed to a manageable number for empirical evaluation. Only variables that are believed to considerably impact user performance, acceptance, and system usability are included for further study in the experimental methodology section of the dissertation.

Finally, a series of decision trees were developed as part of this research effort to provide designers and engineers of in-vehicle systems with a systematic methodology for determining whether a given task function should be performed using manual controls or using speech input. Each decision node and action point on the trees represent various phases of the decision making process, leading to the final result of task-function input-modality allocation. Both current and near-future automotive systems can be evaluated for their utility of being controlled via speech input using this decision tool. Decision criteria such as safety, usability, and driver preference/acceptance provide the logical basis for determining whether a given task-function should be considered for control using speech input or some other conventional input method.

Problem Statement

The goal of the research is to determine the optimal use of automatic speech recognition technology for controlling secondary in-vehicle tasks. The research generates design guidelines that allow designers and engineers to identify which in-vehicle tasks, when performed while concurrently driving an automobile, are amenable to speech recognition technology. The decision tool for determining whether a given task function should be performed using ASR technology or current manual input methods is included in the design guidelines. The design guidelines also include recommendations for ASR system configurations (i.e., system feedback and error correction method) and ASR system performance (i.e., recognition accuracy and recognition errors allowed). These recommendations are based on the results from two experiments that considered aspects of driving performance, usability, and driver preference/acceptance as affected by the ASR system parameters manipulated.

Two experiments were performed instead of a single experiment primarily because of time and cost constraints associated with conducting research on large data spaces. Williges and Williges (1989) explored the problem of human factors design involving large data spaces. These large data spaces contain so many experimental factors that a single experiment evaluating the effects of all the factors would be unreasonable. Concerns such as cost and time involved in collecting data prohibit such traditional experimental approaches. The primary objective of this research strategy is to evaluate the main effects and second-order interactions describing the functional relationship between the independent variables manipulated. Third-order interactions are often difficult to interpret, and second-order approximations are adequate for describing most human performance data (Williges, 1981).

Williges and Williges (1989) developed an integrated research strategy for conducting complex human factors experimentation. The research strategy is outlined in three stages. The first stage is to select the important independent variables that describe the experimental space in question. Once a set of independent variables has been selected, the second stage is to describe the effects of the independent variables through a series of small experiments. It is during this stage that traditional experimental approaches, such as factorial, fractional-factorial, and central composite designs, are employed. Empirical models using polynomial regression are used to describe the functional relationships between the independent variables. By using polynomial regression analysis, results from several studies can be combined into one complex regression equation with some additional data bridging. Following this approach, researchers can form an integrated database of information that addresses the entire set of independent variables of interest. Finally, the third stage of the Williges and Williges research strategy is the optimization of the independent variables. Methods such as response surface methodology, random selection, and partial derivatives are used to optimize the independent variables in question.

The following paragraphs provide justification for the breakdown of the two experiments that evaluated the effects of various ASR system parameters on driving performance, task-function usability, and driver preference/acceptance. However, the results from each experiment were not combined into an integrated database from which empirical models could be developed. Instead, the results from the first experiment were used to establish "optimal" criteria (i.e., recommendations for ASR system design) that were used in the second experiment. Factorial experimental designs were used to evaluate the independent variables for the two experiments.

Experiment 1 used a full-factorial design to investigate three discrete independent variables: recognition errors, error correction method, and ASR system feedback modality. The independent variable, error correction method, is an important design parameter for ASR systems that may be used in automobiles. This experiment determined which error correction method should be utilized by ASR systems controlling secondary in-vehicle tasks. The independent variable, ASR system feedback, is another important design parameter for ASR systems that may be used in automobiles. Experiment 1 also determined the sensory modality in which ASR feedback should be presented. Interactions were expected between recognition errors and error correction method, ASR system feedback and recognition errors, and ASR system feedback and the type of in-vehicle task performed (see Results for Experiment 1 in the Data Analysis section).

The second experiment used a mixed-factorial design to investigate four independent variables: age, input modality, recognition accuracy, and recognition errors. At this stage, optimal methods for displaying ASR system feedback and for correcting ASR system errors were determined from the results of Experiment 1. This "optimized" ASR system was used in Experiment 2 to determine how different levels of age, input modality, recognition accuracy, and recognition errors affected the dependent measures of interest, and at what recognition accuracy level does speech input compare with current manual input methods when performing the in-vehicle tasks.

Economy in data collection is exhibited when comparing the number of data observations required by the proposed sequence of two experiments with a single factorial design using all six independent variables. A single mixed-factorial experiment using all six independent variables of interest would require 74 unique observations. The value of 74 observations results from adding (2 x 2 x 2 x 3 x 3) + 2 (the resultant of 2 age groups, 2 recognition error types, 2 error correction methods, 3 levels of ASR system feedback, 3 levels of recognition accuracy, plus a manual-control input condition and a 100% recognition accuracy speech-input condition). The proposed sequence of two experiments requires only 32 unique observations. The value of 32 observations for the proposed sequence of experiments results from adding (2 x 3 x 3) + (2 x 2 x 3) + 2.

Research Questions

The empirical research strategy addresses ASR system design and performance issues by answering the following questions.

(1) What effect do recognition errors have on driving performance, task-function usability, and driver preference/acceptance, and should the occurrence of certain error types be limited through ASR system design?

(2) What effect does error correction method have on driving performance, task-function usability, and driver preference/acceptance, and is one method most appropriate for improving the interaction between the driver and the ASR system?

(3) What effect does ASR system feedback have on driving performance, task-function usability, and driver preference/acceptance, and is one feedback type (i.e., visual-only, auditory-only, or visual plus auditory feedback) most appropriate for improving the interaction between the driver and the ASR system?

(4) What in-vehicle task-function input modality (manual or speech) is optimal for dual-task performance (i.e., driving while performing a concurrent in-vehicle task), and at what level of recognition accuracy is driving performance, task-function usability, and driver preference/acceptance comparable to manual control of in-vehicle tasks or better?

(5) What in-vehicle task-function input modality (manual or speech) is optimal for dual-task performance (i.e., driving while performing a concurrent in-vehicle task) as a function of driver age, recognition accuracy, and recognition error type?

(6) How appropriate is the proposed decision tool for determining whether a given task-function should be performed manually or using speech input?

EXPERIMENTAL METHODOLOGY

Overview

The objectives of the empirical research were to evaluate the effects of ASR system performance parameters (e.g., recognition errors, error correction method, ASR system feedback, and recognition accuracy), in-vehicle secondary-task-function input modality (manual and speech), and age (younger drivers and older drivers) on driving performance, task-function usability, and driver preference/acceptance. The two experiments measured the eye-movement behavior, steering input behavior, speed maintenance behavior, reaction time to forward scene event, task completion time, and task completion errors of drivers performing various in-vehicle secondary task-functions while concurrently driving.

The experimental tasks consisted of driving an instrumented research vehicle and concurrently performing various in-vehicle tasks. These tasks were determined from the decision tree analysis and included such common task functions as changing radio stations and volume levels; changing climate control modes, fan speed, and temperature; operating the driver information computer; and operating a cellular phone. The tasks were performed using both current manual-control input methods and speech input. A "simulated" speech recognition system was developed and implemented into the research vehicle to control secondary-task functions for on-the-road data collection.

Two experiments were performed to evaluate the effects of ASR system parameters, input modality, and age on driving performance, task-function usability, and driver preference/acceptance of task-function design. Each experiment evaluated a combination of ASR system parameters (i.e., recognition errors, error correction methods, and ASR system feedback), or age, input modality, recognition accuracy, and recognition errors. The decision trees were used to determine task functions that were both amenable and not amenable to ASR technology. After the decision tree analysis was completed, a subset of task functions that were determined amenable and a subset of task functions that were determined not amenable to speech control were evaluated in the following two experiments (see Appendix B). It is important to remember that the decision trees were used in determining whether an in-vehicle task function should be performed using speech input or manual input. Once a task function was recommended for ASR control, the results from the empirical research were used to determine the appropriate ASR system parameters (i.e., system feedback and error correction method) and the appropriate levels of ASR system performance (i.e., recognition accuracy and recognition errors allowed) required before implementing ASR technology inside the automobile to control that function.

The dependent measures recorded, the experimental procedures performed, and the experimental apparatus employed are essentially identical for the two experiments. Only the number and age of subjects, the experimental design, and the manner in which the secondary tasks were performed by the subjects are significantly different for each of the two experiments. Therefore, the dependent variables, the experimental tasks, the equipment and software, and the experimental protocol common to each experiment are described first, followed by a description of the differences between Experiments 1 and 2.

Dependent Variables

The dependent measures used to evaluate driving performance are eye movement behavior (i.e., percent eyes-off-road), steering input behavior, speed maintenance behavior, and reaction time to a forward scene event. These measures were recorded when subjects were driving and concurrently performing an in-vehicle task function during an experimental trial. The measures provided information about the effect the independent variables had on driving while performing an in-vehicle task.

The dependent measures used to evaluate in-vehicle task-function usability are the task completion time and task completion errors when subjects were driving and concurrently performing an in-vehicle task function during an experimental trial. These measures provided information about the effect the independent variables had on task-function usability when performing an in-vehicle task.

The dependent measures used to evaluate driver preference/acceptance of in-vehicle task functions are the results from the subjective questionnaires. The subjective questionnaires were administered immediately after subjects performed an in-vehicle task function and provided information about the effect the independent variables may have on driver preference/acceptance of automotive secondary-control system design. Retrospective subjective opinions were recorded in lieu of concurrent opinions. Measures of driving performance and task-function usability would have been adversely affected if subjects reported opinion data while concurrently performing in-vehicle tasks, and studies have shown that retrospective opinion data does not differ from concurrent opinion in terms of accuracy of information gathered (Ericsson and Simon, 1984). Subjective ratings of overall secondary task characteristics (e.g., input modality, ASR system performance, ease of use, etc.) were measured using five-point and seven-point rating scales that were developed.

Experimental Tasks

The experiments had subjects performing two concurrent tasks. The primary task was an "on-road" driving task. The secondary task consisted of performing a series of ASR-controlled, or manual-controlled, in-vehicle tasks. Primary driving task performance and secondary in-vehicle task performance were expected to vary as the independent variables were manipulated.

All experimental trials included the primary driving task. Subjects controlled the driving task via steering wheel and accelerator/brake-pedal inputs. Driving was performed along a stretch of U.S. Highway 460 between Blacksburg, Virginia and Princeton, West Virginia. This roadway is a four-lane route with numerous intersecting roadways. Cross traffic is relatively low for this section of U.S. Route 460. The 42 mile route between Blacksburg and Princeton is divided for the majority of its length, with two short sections (less than 3 miles each) where the road is not divided as it passes through towns. The posted speed limit is 55 miles per hour; however, the short four-lane undivided sections have posted speeds of 45 miles per hour or 40 miles going through town. The curvature of the roadway and the amount of grade varies substantially for this route as it traverses through the Appalachian Mountains. In general, the roadway has good visibility and low-to-moderate traffic density. Experimental runs occurred only in weather that did not adversely affect visibility or roadway traction. Experimental trials were presented while driving in both directions between Blacksburg and Princeton and an effort was made to have subjects perform in-vehicle tasks on straight road sections with no conflicting traffic.

Subjects were instructed to drive in the right lane at all times, maintain posted speeds, and observe all traffic regulations (e.g., turn restrictions, traffic lights, regulatory and warning signs, etc.). Steering input behavior and speed maintenance behavior were continuously recorded by the data collection computer during execution of in-vehicle secondary tasks and while just normally driving.

A red light, mounted on the hood of the vehicle, was illuminated at random intervals either when the subject was concurrently performing an in-vehicle secondary task or when the subject was just driving. The light was illuminated for 25 percent of the in-vehicle task trials. When the light was illuminated, subjects were instructed to gently depress the brake pedal with their right foot as quickly as possible. Reaction time to the onset of the light was measured from when the light was first illuminated until the subject depressed ("tapped") the brake pedal.

The purpose for including a reaction time to forward-scene event, or red-light, task was to measure the situational awareness of the driver when he or she was driving and performing various in-vehicle tasks versus the driver's situational awareness when only driving. Kames (1979) used a similar task to evaluate driver awareness to events occurring outside the vehicle when the driver was performing various in-vehicle telephone tasks. This reaction-time task provided another measure of workload imposed on the driver when performing in-vehicle tasks using speech input or manual input to accomplish the task. The hypothesis was that reaction times to the onset of the light would be significantly shorter for speech input tasks than for manual input tasks. Therefore, it might be concluded that an ASR-equipped vehicle could reduce workload, and increase situational awareness outside the vehicle, for the driver when he or she is concurrently performing an in-vehicle task while driving.

The secondary task consisted of a series of in-vehicle tasks performed by the subject while concurrently driving. Secondary tasks varied for each experimental trial. These tasks were similar to in-vehicle control tasks performed by the driver while concurrently driving a vehicle. Examples of such tasks are changing the radio station, adjusting the temperature of the climate control system, and dialing a telephone number on a cellular phone (see Table 3 for a list of all the in-vehicle tasks performed by subjects while driving). The specific in-vehicle tasks selected for testing were determined by the decision tree analysis. Task completion time and task execution errors were recorded by the data collection computer and video recording equipment.

An example experimental data collection session proceeded as follows: (1) The subject performed the primary driving task. (2) A secondary task scenario was presented to the subject auditorily by the experimenter. (3) After the subject was instructed to "begin," the subject executed the secondary task while concurrently performing the primary driving task. During this time period, data was collected on driving performance and task-function usability. (4) When the secondary task was completed, driving performance and task-function usability data collection ended and the subject was asked questions to asses driver preference/acceptance of performing the particular secondary task. (5) After the subject finished answering the questions, a new secondary task scenario was presented. (6) After all the secondary tasks for a given experimental session were completed, the subject returned to the starting point of the drive.

Table 3. List of in-vehicle tasks performed by subjects in Experiments 1 and 2.

(1)	Unlock and lock vehicle's doors
(2)	Activate cruise control and set cruise speed to current speed of travel
(3)	Resume cruise speed and then deactivate cruise control
(4)	Activate turn signals, change lanes, and deactivate turn signals
(5)	Activate and deactivate windshield wipers
(6)	Adjust temperature control on the climate system
(7)	Adjust fan speed on the climate system
(8)	Change climate mode (e.g., panel vents, defrost, etc.) on the climate system
(9)	Display average fuel economy and reset it on the driver information computer
(10)	Display trip distance on the driver information computer
(11)	Tune radio to station stored in preset (or memory)
(12)	Tune radio to specific radio frequency
(13)	Seek “up” to next receivable radio station using seek function on the radio
(14)	Mute the volume on the radio
(15)	Dial a number stored in memory using the cell phone
(16)	Dial home telephone number using the cell phone

Equipment and Software

The research vehicle used in the experiments was a 1995 Oldsmobile Aurora with an automatic transmission, analog instrument cluster, and electronic audio, climate-control, and driver-information systems. Additionally, the vehicle was equipped with a seven-inch diagonal flat-panel LCD display mounted to the right of the speedometer on the instrument cluster. The LCD display served as the ASR system display and provided visual feedback to the driver during ASR task execution. The speech-operated cell-phone tasks also employed this display. To minimize risk for the subject, the vehicle was equipped with a fire extinguisher, a first-aid kit, and a cellular telephone which could be used to contact public safety in the event of an emergency.

The vehicle contained two types of control systems for performing various in-vehicle task functions. The first control system comprised the conventional manual controls used to perform various in-vehicle tasks. The second control system comprised a "simulated" speech recognizer used to control various in-vehicle task functions. Subjects spoke appropriate commands to execute task functions. For example, a subject would say "climate--mode--defrost" and the climate control system's "defrost" mode would be activated. A second experimenter in the vehicle acted as the speech recognition system and executed subjects' spoken commands using a computer and customized "controller-box" to operate secondary task functions. A "Wizard of Oz" methodology (Gould et al., 1983) was employed to make subjects believe that the research vehicle was equipped with an actual speech recognition system. All subjects tested were unaware that spoken commands were actually executed by the experimenter and not a real speech recognition system. A simulated speech recognition system was used so that system parameters such as recognition errors, system feedback, error correction methods, and recognition accuracy could be manipulated during the experiment.

The vehicle was also equipped with four CCD video cameras mounted inside the research vehicle. The first camera, positioned inside the center rearview mirror, was used to record drivers' eye movements. The video image recorded was of the driver's head with some additional space around the head to accommodate any head-movements by the driver during data collection. A second camera, positioned on the roof-liner in the center of the vehicle, was used to record the instrument panel where the driver interacted with certain in-vehicle manual controls used in the study. The second camera's video image recorded the center console where the audio, climate-control, and driver-information systems are located. A third camera, positioned behind the center rearview mirror, was used to record the forward driving scene. The fourth camera, mounted inside the driver's-side (left-outside) mirror, was used to record lane position of the vehicle with respect to the centerline of the road. The video image from each camera was combined using a multiplexer, and the four images were recorded simultaneously in a quad-split view on a video cassette recorder. A time code generator was used to insert a digital time stamp on each frame of the videotape recording. Microphones were mounted inside the vehicle so all commands and comments made by the subjects and the experimenter could be recorded onto the video tape.

Two computers were mounted inside the vehicle. The first computer, referred to as the ASR computer, controlled the speech-input secondary tasks (i.e., the in-vehicle secondary tasks performed by subjects using speech commands). The second computer, referred to as the data collection computer, monitored and recorded data from the sensors in the vehicle used to measure steering behavior and speed maintenance behavior. The data collection computer was also used to control and record the forward-scene reaction time event and the task completion timer for in-vehicle secondary tasks.

Software, developed by the Virginia Tech Center for Transportation Research, was used to control the speech-input secondary tasks and to record all relevant driving performance data (e.g., steering input behavior, speed maintenance behavior, reaction-time to forward scene event, and task completion time).

Experimental Protocol

After arriving at the Virginia Tech Center for Transportation Research, each subject was asked to read and sign an informed consent form (see Appendices C and D for Experiments 1 and 2, respectively). After reading the informed consent, the experimenter checked for a valid driver's license; administered the health screening questionnaire (see Appendix E); and reviewed the procedures to the subject, answering any questions from the subject.

The experimenter then led the subject to the research vehicle and demonstrated the various displays and controls of the vehicle to the subject. Next, the experimenter trained the subject on the various in-vehicle tasks that were to be performed during the study. The experimenter demonstrated and explained each in-vehicle task to be performed by the subject during the experimental trials. The subject was provided adequate training in proper execution of the tasks, as well as adequate practice. Upon completion of the demonstrations, the subject's questions were answered. Only when the subject was comfortable with the experimental apparatus and procedures was he or she asked to begin driving the vehicle. Once on the road, and the subject was comfortable driving the vehicle and performing the tasks, the experimental data collection session began.

Prior to each experimental treatment-blocking condition, subjects were instructed on the appropriate task characteristics for that session (e.g., ASR system feedback provided, error correction method required, input modality, etc.). During each experimental trial measures of driving performance and task-function usability were recorded. After completion of an in-vehicle task, the subject was asked a series of five-point (or seven-point) subjective rating scale questions to assess driver preference/acceptance of the task-function design (see Appendices F and G for Experiments 1 and 2, respectively). After answering the questions, a brief rest period (approximately one minute in length) was provided, and the procedure repeated for the next treatment condition. During rest periods, eye movement behavior, steering input behavior, speed maintenance behavior, and reaction time to the forward scene event were measured to obtain a baseline value of driving performance from which driving performance while concurrently performing in-vehicle tasks is compared.

Forty-eight in-vehicle secondary tasks were performed during each of the experimental treatment-blocking sessions. After the last in-vehicle task was completed--on the last blocking session of the experiment--the subject was debriefed and paid for his or her participation.

Experiment 1

Subjects. Five male and five female subjects, ages 20 to 26 years (mean age = 21.1 years), participated in the experiment. All drivers presented a valid driver's license before being allowed to participate. Each subject was tested individually in three separate experimental sessions which lasted approximately two hours each. All subjects were paid $5.00 per hour for their participation.

Experimental design. The experiment used a 2 x 3 x 3 within-subjects factorial design. The three independent variables for this experiment were error correction method, recognition errors, and ASR system feedback modality. The within-subjects factor--error correction method--had two levels: (1) "cancel" and repeat command and (2) simply repeat command. The within-subjects factor--recognition errors--had three levels: (1) no error, (2) rejection error, and (3) substitution error. Insertion errors were eliminated as a level of interest because of their infrequency of occurrence in automotive ASR systems, which employ a "push-to-talk" switch to activate the system. The within-subjects factor--ASR system feedback modality--had three levels: (1) visual-only feedback, (2) auditory-only feedback, and (3) a combined auditory plus visual feedback.

The order of presentation for in-vehicle task functions was grouped by ASR system feedback modality. Only one level of ASR system feedback modality was employed for each experimental session. The order of presentation for error correction method was counterbalanced across subjects for each experimental session. The order of presentation for recognition errors was randomized for each session by error correction method segment of data collection. Finally, the order of presentation for experimental sessions was counterbalanced across subjects using a balanced Latin Square design; that is, every session-type (i.e., feedback modality condition) preceded and followed every other session-type an equal number of times.

Task specifics. The secondary tasks for Experiment 1 were accomplished through speech input. Subjects performed identical tasks for all combinations of ASR system parameters. The ASR system feedback modality remained constant for a given experimental session, and feedback was provided at the appropriate times when a subject was performing an in-vehicle task. Subjects used both error correction methods during a single experimental session. Recognition errors occurred at predetermined intervals when a subject was performing an in-vehicle task requiring the subject to employ an error correction method. Subjects were given scenarios to facilitate in-vehicle task execution.

The type of auditory-only feedback provided was either the message "Please repeat," which was played when a rejection error occurred, or an auditory "beep," which was played when a substitution error occurred, or when no error occurred, to indicate that a command was recognized and executed (even if the command recognized was incorrect--a substitution error). The type of visual-only feedback provided was either the message "Please repeat" (when a rejection error occurred), or "Recognized" (when no error occurred or when a substitution error occurred), which was displayed on the LCD display mounted on the instrument panel. The appropriate system display also changed to reflect a recognized command or substitution error (e.g., the radio station display changed to reflect the command for a station change, the climate-control mode display changed to reflect the command for a climate-mode change, etc.). The combined auditory plus visual feedback condition included both the auditory-only and visual-only feedback described above.

Subjects performed each in-vehicle task three times per feedback modality by error correction method blocking session. One time the in-vehicle task had no errors, another time the task contained a rejection error, and the third time the task contained a substitution error. The order in which error types occurred was randomized.

Experiment 2

Subjects. Six subjects--three male and three female--ages 21 to 27 years (mean age = 22.8 years) and six subjects--three male and three female--ages 65 to 78 (mean age = 71.7 years) participated in this experiment. Significant differences in driving performance and behavior were expected between these age classifications. All drivers presented a valid driver's license before being allowed to participate. Each subject was tested individually in three separate experimental sessions which lasted approximately two and one-half hours each. The younger subjects were paid $10.00 per hour for their participation, and the older subjects were paid $15.00 per hour for their participation.

Experimental design. The experiment used a 2 x 2 x 3 mixed-factorial design with two control conditions: (1) a manual-control input condition and (2) a 100% recognition accuracy speech-input condition. The independent variables for the mixed-factorial design were age (the between-subjects variable), recognition error (a within-subjects variable), and recognition accuracy (another within-subjects variable). The within-subjects factor--recognition error--had two levels: (1) rejection errors only and (2) substitution errors only. The within-subjects factor--recognition accuracy--had three levels: (1) 90% recognition accuracy, (2) 75% recognition accuracy, and (3) 60% recognition accuracy.

The between-subjects factor included two levels of age. The two levels of age investigated were: (1) younger drivers (between the ages of 18 and 35 years) and (2) older drivers (at least 65 years of age and retired). In general, research studies have shown that older drivers differ significantly from younger drivers in measures of driving performance (see Allen, Stein, Rosenthal, Ziedman, Torres, and Halati, 1991; Dingus et al., 1989; Dingus, Gellatly, and Reinach, 1997; Dingus, Hulse, Krage, Szczublewski, and Berry, 1991; Dingus, Hulse, McGehee, and Manakkal, 1994; Kiefer and Gellatly, 1996; Kurokawa, 1990; Marin-Lamellet and Dejeammes, 1995; Mollenhauer, Lee, Cho, Hulse, and Dingus, 1994; Paelke, 1993; Perez and Mast, 1992; Serafin et al., 1993; Sivak, Flannagan, and Gellatly, 1991; Snyder and Monty, 1986; Walker, Alicandri, Sedney, and Roberts, 1991; Wierwille, 1993).

The number of drivers over the age of 55 is increasing (Ng and Barfield, 1993). Visual acuity, reaction time, and cognitive skills all tend to worsen with age. Introducing new technologies such as automatic speech recognition could have a huge impact on driving performance for this segment of the driving population. The majority of automotive advanced-technologies are first introduced in "high-end" vehicles (i.e., high-priced luxury cars), which are primarily purchased by the older driving population. The reason for this fact is that introducing new technologies can significantly increase the cost of the vehicle. This cost increase is more easily justified on high-priced vehicles than on lower-priced vehicles. Older people generally have more income than younger people and are therefore able to afford the higher cost of a vehicle equipped with such advanced technologies. Therefore, including age as an independent variable for Experiment 2 was critical for research related to both driving safety and performance, as well as the usability and acceptance of new technologies introduced into the automobile.

The order of presentation for in-vehicle task functions was grouped by experimental session. During the first experimental session, subjects experienced the two control conditions: the manual-input condition and the 100% recognition accuracy speech-input condition. Order of presentation for control conditions was counterbalanced across subjects. Subjects experienced one of the two recognition error conditions for the final two experimental conditions. Order of presentation for the recognition error condition was counterbalanced across subjects. During each of the final two experimental sessions, subjects experienced all three levels of recognition accuracy. The order of presentation for recognition accuracy level was grouped by recognition error type and counterbalanced across subjects using a balanced Latin Square design; that is, every recognition accuracy level preceded and followed every other recognition accuracy level an equal number of times.

Task specifics. The secondary tasks for Experiment 2 were accomplished through speech input and through manual input. Subjects performed identical tasks using both input modalities. However, subjects used only speech input during experimental sessions two and three. Task-function input modality remained constant for a given treatment-blocking session. Subjects were given scenarios to facilitate in-vehicle task execution.

Subjects performed each in-vehicle task three times per treatment-blocking condition (e.g., three times per control condition, or three times per recognition error by recognition accuracy condition). Recognition accuracy levels were achieved by including an appropriate number of recognition errors out of the total command input sequences, which resulted in the desired accuracy level for a given treatment-blocking condition. For example, there were 114 speech-command inputs required to correctly perform the 16 different in-vehicle tasks a total of three times. Therefore, to achieve a 90% accuracy level, 13 command inputs would have a recognition error occur and 127 inputs would be error-free. The order in which errors were presented was randomized.

During speech input tasks, the combined auditory plus visual feedback was used to inform subjects of the status of their interaction with the ASR system when performing a task. The error correction method used by subjects was the "cancel" and repeat command method. Subjects were instructed to say the command "cancel" and repeat the correct command vocabulary when a substitution error occurred. The levels of ASR system parameters selected as constants for Experiment 2 (i.e., ASR system feedback modality and error correction method) reflected optimal ASR system parameters as they affect driving performance, task-function usability, and driver preference/acceptance, which were determined empirically in Experiment 1.

RESULTS AND DISCUSSION

Overview

Multivariate analyses of variance (MANOVA) were performed on the data to determine the effect changes in the independent variables had on the dependent measures as a group. Separate univariate analyses of variance (ANOVA) were performed on the individual dependent measures to assess their sensitivity to the independent variables whenever significant MANOVA results were found. Student Newman-Keuls post-hoc tests were performed on significant main effects from the univariate analyses. The Statistical Analysis System (SAS) Version 6.11 for Open Systems computer-software program was used to perform these statistical analyses (SAS Institute, 1995).

McNemar change, Chi-square, and Cochran *Q* tests were performed on data recorded for the number of task completion errors by subjects. The results from these nonparametric tests determined what effect the independent variables had on task errors made by drivers. Task error data is categorical thereby requiring the use of nonparametric statistical tests to evaluate how changes in the independent variables affected the number of task completion errors made by subjects.

A separate MANOVA was performed for each data set from Experiments 1 and 2. The use of multivariate statistics was appropriate for this research because the experimental approach required the analysis of numerous dependent variables that were all correlated with one another to some degree. Normally, a number of independent tests would have to be performed on the same sample to gain inference about the effects the independent variables had on the dependent measures recorded. The MANOVA technique controls for the problem of inflated error rate with multiple dependent variables that are related (Tabachnick and Fidell, 1989). The sensitivity of all the dependent measures, as a single group, to changes in the independent variables is tested by performing a MANOVA. Whenever significant results were found from the initial MANOVA procedure, separate univariate analyses of variance were performed for each of the dependent measures to determine the effects of the independent variables on each individual dependent measure.

Simple effects tests were conducted whenever a significant error type by error correction method interaction was found in Experiment 1. For this condition, drivers did not have to employ an error correction method when no error occurred during task performance. When a rejection error occurred, drivers were instructed through ASR system feedback to repeat the command input. Therefore, both the no error and rejection error conditions were not relevant in a significant error type by error correction method interaction. A simple effects test was performed to determine if error correction method had a significant effect on the dependent measures recorded whenever substitution errors occurred.

The criterion for statistical significance chosen for this research was $\alpha = 0.10$. The criterion was relaxed from the conventional $\alpha = 0.05$ because the primary purpose for this research was to develop a comprehensive set of qualitative and quantitative design guidelines for the use of ASR systems in automotive applications. The guidelines and recommendations are based primarily on the results from the empirical research. The relaxed criterion allowed empirical results to be included that showed the effects that ASR system parameters had on measures of driving performance, task-function usability, and driver preference/acceptance. These additional effects were helpful in substantiating the qualitative and quantitative recommendations regarding ASR system designs that are currently being considered for use in automotive applications. However, it is important not to place too great an emphasis on any single significant result because the probability of finding a difference between levels of the independent variable, when no difference actually exists, is increased (e.g., the probability of a Type I error is increased).

Data collection was conducted in two stages; therefore, the results are presented in two stages. The first stage of data collection was performed to evaluate the effect of ASR system feedback modality, recognition error type, and error correction method on the dependent measures recorded. The second stage of data collection evaluated the effect of driver age, input condition, and in-vehicle task type on the same dependent measures recorded in Experiment 1. Further analyses were performed to determine the effect of driver age, recognition accuracy level, and recognition error type on the dependent measures recorded for Experiment 2.

Two statistical tests were performed on largely the same data set for Experiment 2. The first test evaluated the independent variables age, input condition (i.e., manual input and speech input at four levels of recognition accuracy and with two error types predominant), and in-vehicle task type. This test was performed to determine the effect the different input conditions had on measures of driving performance, task-function usability, and driver preference/acceptance of task functions. Comparisons between the different levels of speech-input system performance and the manual-input control condition were made from this analysis. Driver age was included to determine the effects of age on the dependent measures recorded, and to determine if any interactions between age and input condition existed. The results from this analysis were useful in determining levels of ASR system performance that were comparable to or better than manual-input methods in terms of the dependent measures recorded. Results from this analysis also provided a level of validation for the outputs from the decision trees used to determine task-function input-modality allocation. Simple effects tests were performed on all significant task type by input condition interactions to evaluate how well the decisions trees predicted whether a given task function should be performed using speech input or some type of convention manual control input. Tables for the results of the simple effects tests are included in this section.

The second test evaluated the independent variables age, recognition accuracy (i.e., 90%, 75%, and 60%), and recognition error type (i.e., rejection errors and substitution errors). This test was performed to determine the effects of recognition accuracy and error type on the dependent measures recorded, and to determine if any interactions between accuracy and error type existed. Driver age was included in this analysis to determine if age interacted with accuracy or error type (e.g., older drivers may have been less tolerant to substitution errors, in terms of driving performance, than younger drivers). Data from the manual-input condition and the 100% recognition accuracy condition were dropped to perform this analysis.

The criterion for statistical significance was changed for Experiment 2 to $\alpha = 0.05$ to reduce the probability of spuriously significant results when multiple tests are performed on largely the same data. The criterion for statistical significance was determined from the following equation used to determine familywise (FW) Type I error rate when multiple tests are performed on a set of data (Keppel, 1991).

$$\alpha_{FW} = 1 - (1 - \alpha)^{c}$$

α_{FW} = criteria for statistical significance (i.e., familywise Type I error rate);

α = criteria for statistical significance for single test (i.e., per comparison Type I error rate); and

c = number of statistical tests performed.

Experiment 1

Analysis of variance tables are presented in Appendix H for the driving performance data, task-function usability data, and driver preference/acceptance data. Only significant main effects and two-way interactions are reported and discussed in this section. It is important to note that a number of F-ratios were less than 1.0 for the ANOVA results in Experiment 1. F-ratios less than one indicate the likelihood for heterogeneity of variance between sample populations. An assumption of the ANOVA technique is that each of the sample populations of scores has the same variance (Howell, 1992). Although analysis of variance is a very robust statistical procedure, violating the assumption of homogeneity of variance can lead to an inflation in Type I error (Keppel, 1991). The reader is cautioned of this fact; therefore, it is important not to place too great an emphasis on any single significant result because the probability of finding a difference between levels of the independent variable, when no difference actually exists, is increased.

Measures of driving performance. Three major categories of dependent measures were used in evaluating driving performance. The categories include (1) steering input behavior, (2) speed maintenance behavior, and (3) reaction time to a forward scene event. Results are presented for each category when significant main effects or interactions were found. The significance of the results and the implications they may have on ASR system design are also discussed in this section when appropriate.

Steering input behavior was evaluated through measurements of steering variability and the number of steering wheel reversals greater than a specified number of degrees of steering wheel rotation (e.g., 6 and 12 degrees). Steering variability (i.e., variance), while subjects were performing in-vehicle tasks, was analyzed in a full-factorial multivariate analysis of variance (MANOVA) against the independent variables feedback modality, error correction method, and error type. Only one main effect was significant: error type, $F(2,18) = 10.61$, $p < 0.001$. A complete ANOVA table for steering variability is shown in Appendix H, Table H-1.

The effect of error type on steering variability is shown in Figure 2. A Student Newman-Keuls post-hoc test showed that only substitution errors significantly increased steering variance compared with the no error and rejection error conditions. No difference in steering variability was found between the rejection error and no error conditions.

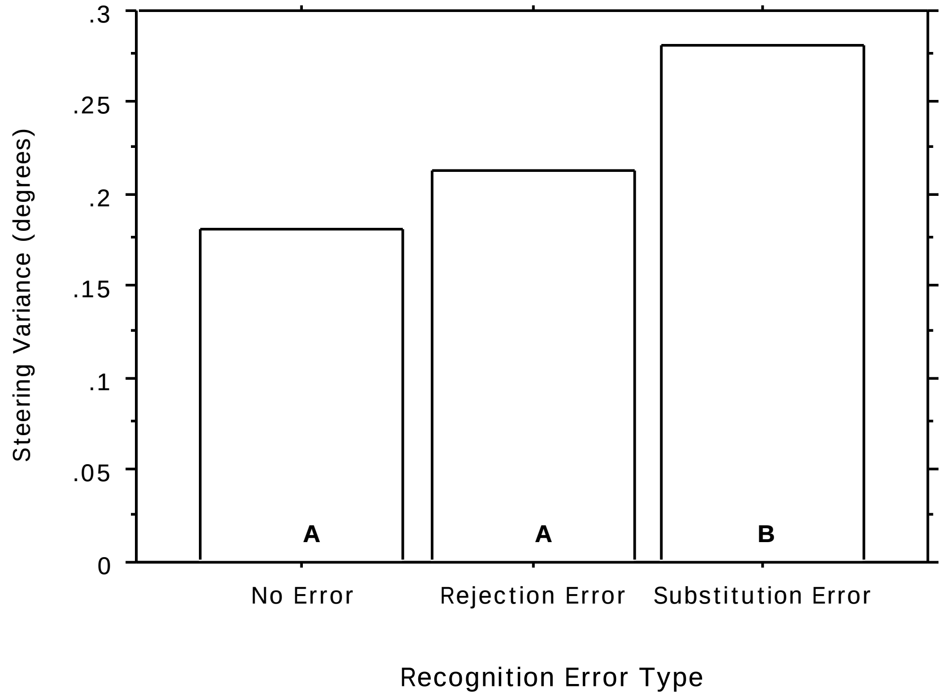

Figure 2. Mean steering variance during task performance as a function of recognition error type. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

The number of steering-wheel-reversals per minute greater than 6 degrees, and the number of steering-wheel-reversals per minute greater than 12 degrees were analyzed in the full-factorial MANOVA against the independent variables feedback modality, error correction method, and error type. No significant effects were found for both "Steering Reversals > 6°" and "Steering Reversals > 12°" analyses. Complete ANOVA tables for steering wheel reversals greater than 6 and 12 degrees are presented in Appendix H, Tables H-2 and H-3.

Speed maintenance behavior was evaluated through measures of accelerator-pedal position variability, mean vehicle velocity (in miles per hour), vehicle velocity variability (in miles per hour), and the number of brake activations per minute. All measures of speed maintenance behavior analyzed were recorded while drivers were performing in-vehicle tasks.

Accelerator position variance was analyzed in the full-factorial MANOVA against the independent variables feedback modality, error correction method, and error type. Only one main effect was significant: error type, $\underline{F}(2,18) = 9.77$, $\underline{p} < 0.005$. A complete ANOVA table for accelerator position variability is shown in Appendix H, Table H-4.

The effect of error type on accelerator position variance is shown in Figure 3. Substitution errors accounted for the greatest variance in accelerator pedal position. A Student Newman-Keuls post-hoc test found no statistical difference between the no error and rejection error conditions. Substitution errors, however, produced significantly larger accelerator position variances compared with both the rejection error and the no error conditions.

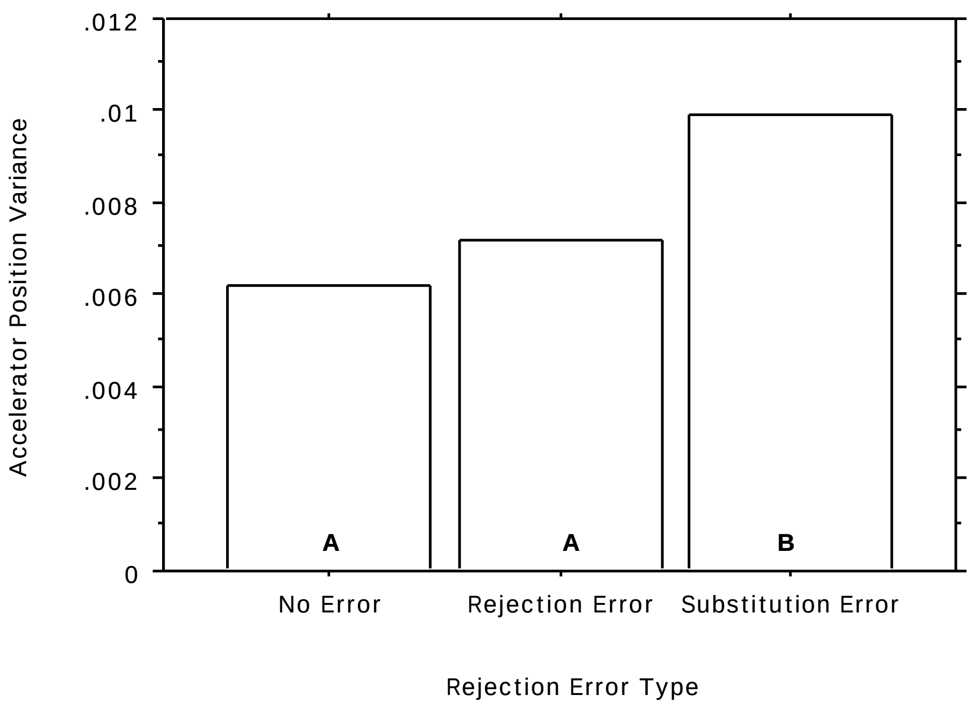

Figure 3. Mean accelerator position variance during task performance as a function of recognition error type. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

Mean vehicle velocity and vehicle velocity variance during task performance were also analyzed in the full-factorial MANOVA against the independent variables feedback modality, error correction method, and error type. No significant effects were found for the mean vehicle velocity measure. However, the main effect of error type was significant ($\underline{F}(2,18) = 19.77$, $\underline{p} < 0.001$) for the vehicle velocity variance measure. Complete ANOVA tables are presented in Appendix H, Tables H-5 and H-6.

The significant error type effect for the vehicle velocity variance measure is shown in Figure 4. Again, it was shown that substitution errors accounted for the greatest variance in a dependent measure when drivers were performing speech-input tasks. The Student Newman-Keuls post-hoc test found no statistical difference between the no error condition (variance = 2.2 mph) and rejection error condition (variance = 2.3 mph). The substitution error condition, however, produced a significantly larger vehicle-speed variance compared with both rejection error and the no error conditions (variance = 5.5 mph).

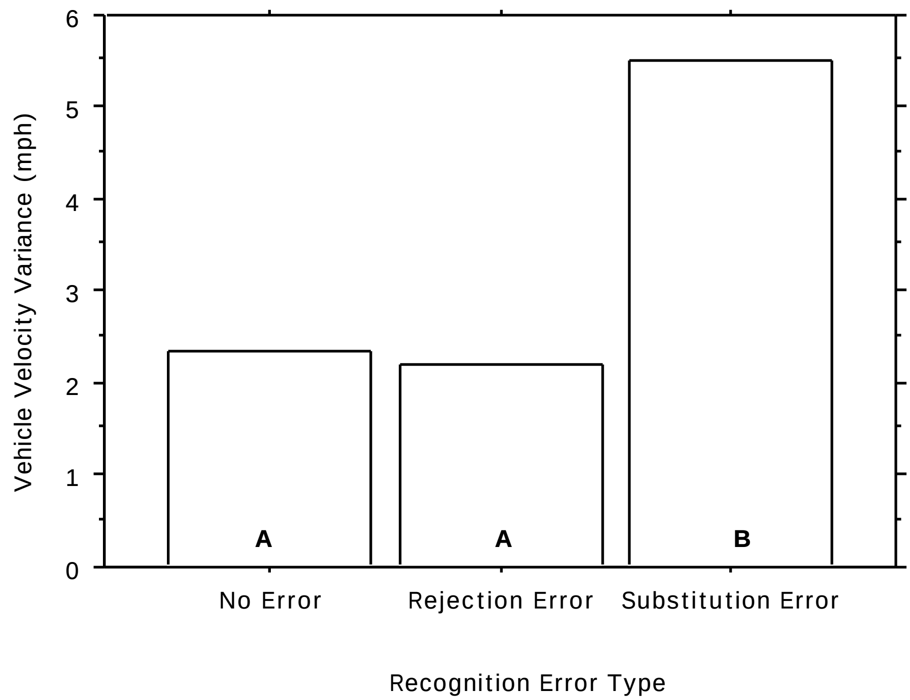

Figure 4. Mean vehicle velocity variance during task performance as a function of recognition error type. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

Finally, the number of brake activations per minute was analyzed in the full-factorial MANOVA against the independent variables feedback modality, error correction method, and error type. Analysis of brake activations provides further information about driving performance when subjects were driving and concurrently performing in-vehicle tasks under different ASR system design parameters. Two effects were found from the analysis: (1) an error type main effect, $\underline{F}(2,18) = 9.54$, $\underline{p} < 0.005$; and (2) a feedback by error type interaction, $\underline{F}(4,36) = 2.17$, $\underline{p} < 0.100$. A complete ANOVA table for brake activations is shown in Appendix H, Table H-7.

The error type effect is shown in Figure 5. The substitution error condition result in the greatest number of brake activations per minute when performing speech-input tasks. A Student Newman-Keuls post-hoc test revealed no statistical difference between the rejection error condition and the no error condition. However, the substitution error condition resulted in a significantly larger number of brake activations per minute than either the no error condition or the rejection error condition when drivers were performing speech-input tasks.

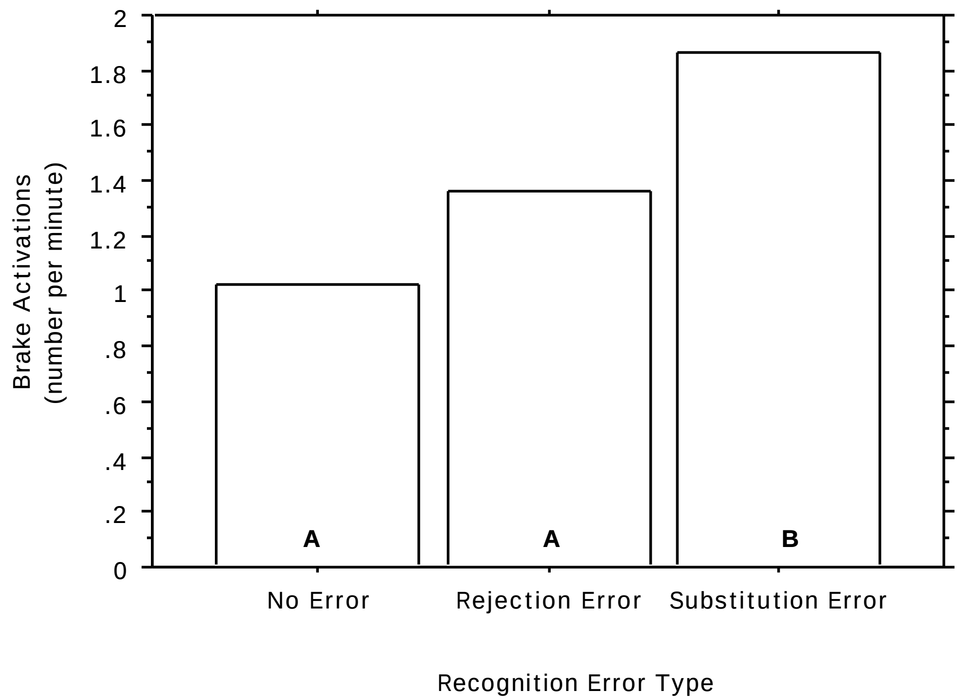

Figure 5. Mean number of brake activations per minute during task performance as a function of recognition error type. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

The feedback modality by error type interaction is shown in Figure 6. Visual plus auditory feedback resulted in the lowest number of brake activations per minute, while visual only feedback resulted in the highest number of brake activations per minute for both the rejection error condition and the substitution error condition. However, the opposite was found for the no error condition; visual plus auditory feedback resulted in the highest number of brake activations per minute, while visual only feedback resulted in the lowest number of brake activations per minute. The auditory only feedback condition resulted in a number of brake activations that was between the number of activations for the other two feedback modality conditions across all recognition error types.

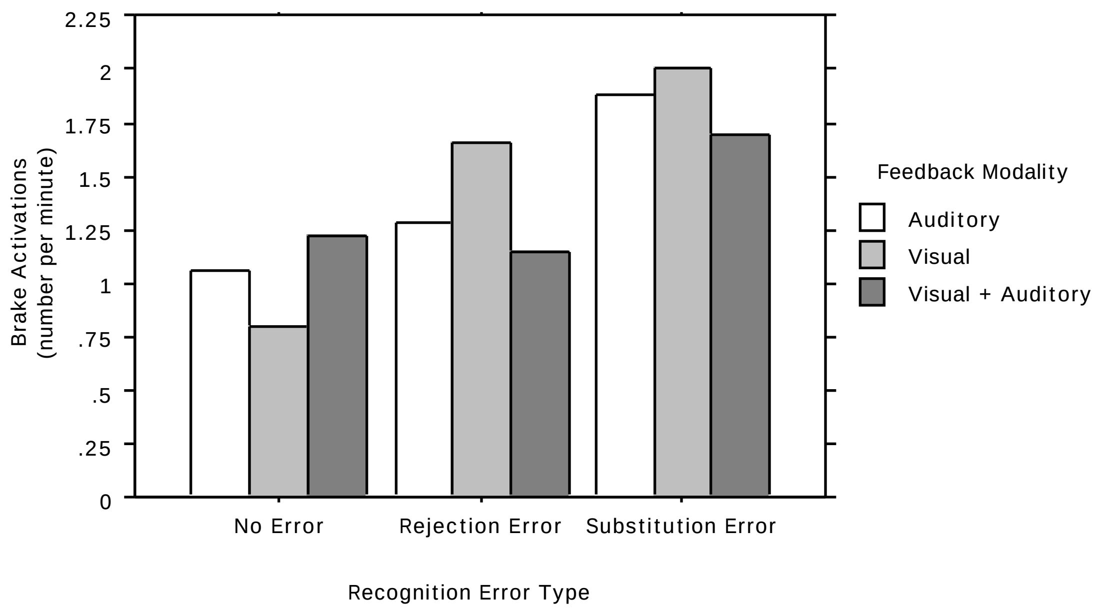

Figure 6. Mean number of brake activations per minute during task performance as a function of ASR system feedback modality and recognition error type.

Reaction time to a forward scene event refers to drivers' brake response time to the light mounted on the hood of the vehicle (see Experimental Methodology). This task simulated an unexpected event to which drivers were instructed to respond with a brake input. Response time was analyzed in two separate full-factorial ANOVAs: (1) against the independent variables gender and driving segment when the light was illuminated (either during task performance, or during driving-only), and (2) against the independent variables feedback modality, error correction method, and error type. No significant effects or interactions were found for either analysis (see Appendix H, Tables H-8 and H-9).

A possible explanation for the lack of significant reaction-time differences in the gender by driving segment analysis is that speech-input tasks may not have drawn a significant amount of drivers' attention away from the forward road scene. Drivers were able to respond to the light when performing speech-input tasks as quickly as they could when just driving the vehicle. Attention used in performing ASR tasks was most likely drawn from "spare" resource capacity.

Variations in ASR system parameters had no effect on reaction time when drivers were performing speech-input tasks (see Appendix H-9). None of the ASR design parameters tested demanded a significant amount of attentional resources. If a design parameter (e.g., the type of system feedback) had drawn more attentional resources from drivers, a statistically significant difference in reaction time would have resulted for the feedback modality by error correction method by error type analysis.

Task-function usability. Two dependent measures were used in evaluating the usability of the various in-vehicle tasks as performed under the various ASR system parameter conditions (i.e., the independent variables tested). The two measures were (1) task completion time and (2) task completion errors. It was hypothesized that drivers experience increased task-function usability with shorter task completion times and/or fewer task errors. Results are presented for each measure when significant main effects or interactions were found. The significance of the results and the implications they may have on ASR system design are also discussed in this section when appropriate.

Task completion time refers to the time duration--in seconds--that it took drivers to complete an in-vehicle task. Time to complete speech-input tasks was analyzed in the full-factorial MANOVA against the independent variables feedback modality, error correction method, and error type. The main effect, error type, was significant ($F(2,18) = 20.75$, $p < 0.001$), as was the error correction method by error type interaction ($F(2,18) = 8.21$, $p < 0.005$). A complete ANOVA table for task completion time is shown in Appendix H, Table H-10.

The error type main effect and error correction method by error type interaction are shown in Figures 7 and 8, respectively. A Student Newman-Keuls post-hoc test for the error type main effect found no statistical difference between the rejection error and substitution error conditions. The no error condition, however, resulted in significantly shorter task completion times compared with both the rejection error and substitution error conditions (see Figure 7).

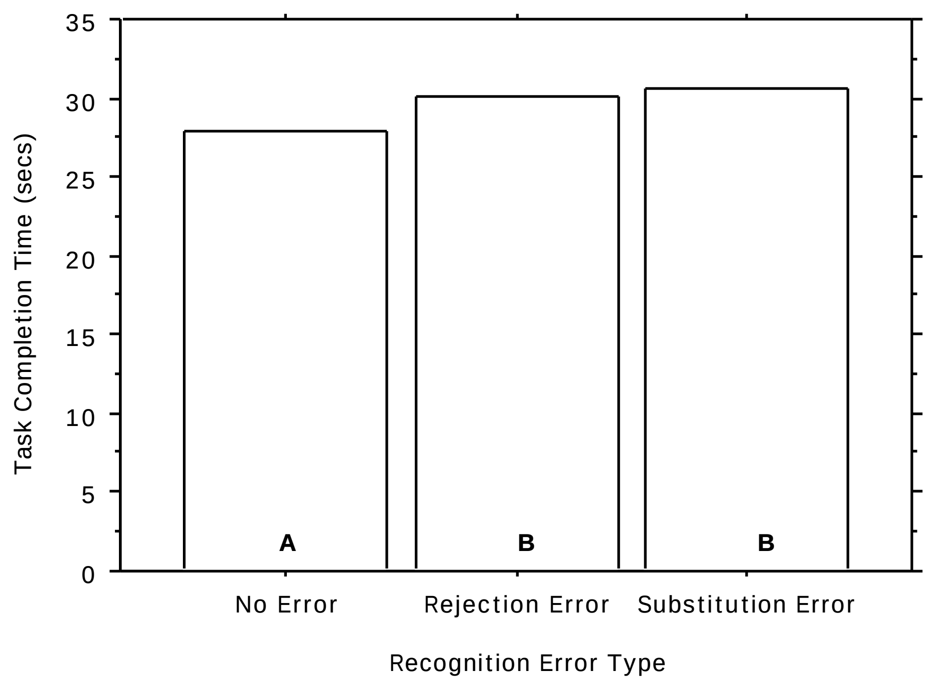

Figure 7. Mean task completion time as a function of recognition error type. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

The error correction method by error type interaction (see Figure 8) shows that the "Simply Repeat" error correction method condition resulted in shorter task completion times than the "Cancel and Repeat" error correction method when substitution errors occurred. Drivers did not have to employ an error correction method under the no error condition. Under the rejection error condition, drivers were instructed through system feedback to repeat the command input. Therefore, both the no error and rejection error conditions are not relevant for this interaction. A simple effects test found an effect of error correction method on task completion time when substitution errors occurred ($\underline{F}(1,9) = 3.87$, $\underline{p} < 0.100$). Drivers were able to complete tasks in a shorter period of time when a substitution error occurred using the "Simply Repeat" error correction method than using the "Cancel and Repeat" error correction method.

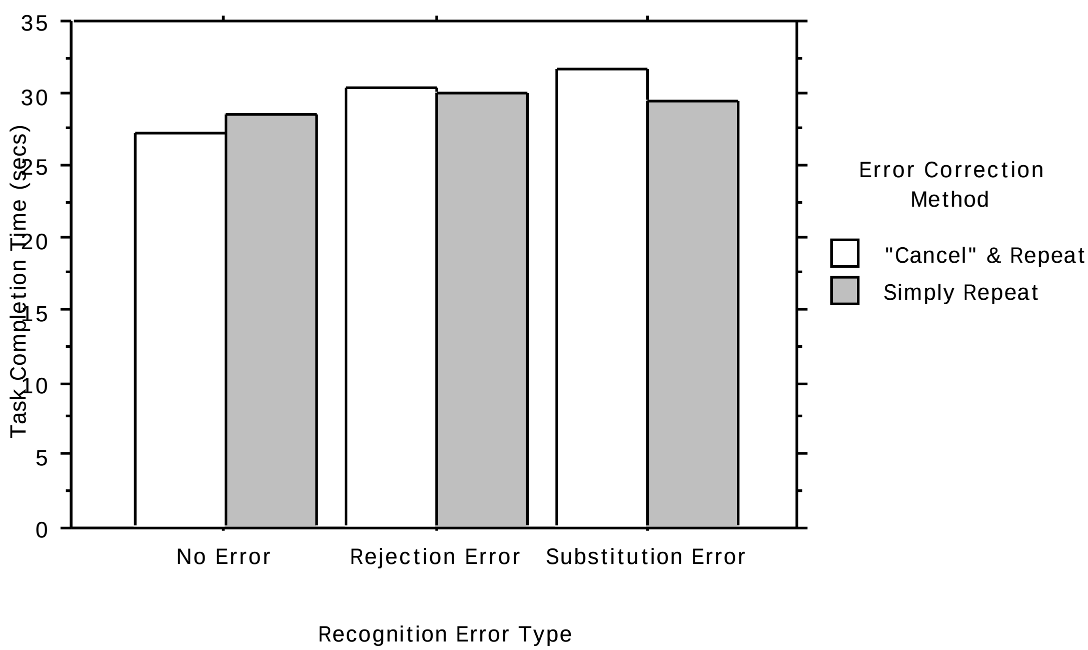

Figure 8. Mean task completion time as a function of recognition error type and error correction method.

Task completion errors were recorded whenever drivers did not correctly complete a speech-input task. To analyze task errors, either a McNemar change test or a Cochran *Q* test (Siegel and Castellan, 1988) was performed to examine the effects of feedback modality, error correction method, error type, and type of in-vehicle task on correct task completion by drivers. The McNemar change test was performed to test the error correction method independent variable, and the Cochran *Q* test was performed to test the feedback modality, the error type, and the task-type independent variables. The type of test used was determined by the number of levels for the independent variable examined. These nonparametric tests determined the significance of differences between two or more related groups (i.e., the levels of the independent variables) with respect to the number of task errors that occurred. It was appropriate to use nonparametric statistics for this data because it was only categorical-type data.

The results showed no effect of feedback modality (Q (2,$\underline{N}$=2880) = 0.041, $\underline{p} < 0.980$) or error correction method (X^2(1,$\underline{N}$=2880) = 0.022, $\underline{p} < 0.900$) on task completion errors. However, correct completion of speech-input tasks was affected by error type (Q (2,$\underline{N}$=2880) = 42.897, $\underline{p} < 0.001$) and the type of task being performed (Q (15,$\underline{N}$=2880) = 1943.826, $\underline{p} < 0.001$). Substitution errors accounted for the significant proportion of task errors recorded (180 errors out of a total 182 recorded errors). The turn signal task, the windshield wiper task, the "seek-up" the radio stations task, and the cellular phone preset number task accounted for the significant proportion of task errors recorded (34, 58, 34, and 37 errors out of a total 182 recorded errors).

Driver preference/acceptance of task functions. Five-point Likert-type subjective questionnaires (see Appendix F) were used to evaluate driver preference/acceptance of in-vehicle task functions as affected by ASR system feedback modality, error correction method, and recognition error type. The questionnaires measured drivers' perceived "ease-of-use," "comfort level," and "distraction level" experienced while performing the in-vehicle tasks. Results are presented for each measure when significant main effects or interactions were found. The significance of the results and the implications they may have on ASR system design are presented in this section when appropriate.

The first question a driver was asked after completing a task was, "Using the five-point ease-of-use scale, how easy was it to perform this task while driving the vehicle?" The results from the "ease-of-use" question were analyzed in the full-factorial MANOVA against the independent variables feedback modality, error correction method, and error type. The main effect, error type, was significant ($\underline{F}(2,18) = 20.62$, $\underline{p} < 0.001$) as was the error correction by error type interaction ($\underline{F}(2,18) = 3.91$, $\underline{p} < 0.050$). A complete ANOVA table for ease-of-use rating is shown in Appendix H, Table H-11.

The significant error type effect and error correction method by error type interaction are shown in Figures 9 and 10, respectively. A Student Newman-Keuls post-hoc test for the error type main effect found statistical differences between all three error type conditions. The no error condition resulted in the highest ease-of-use rating, followed by the rejection error condition, and then by the substitution error condition.

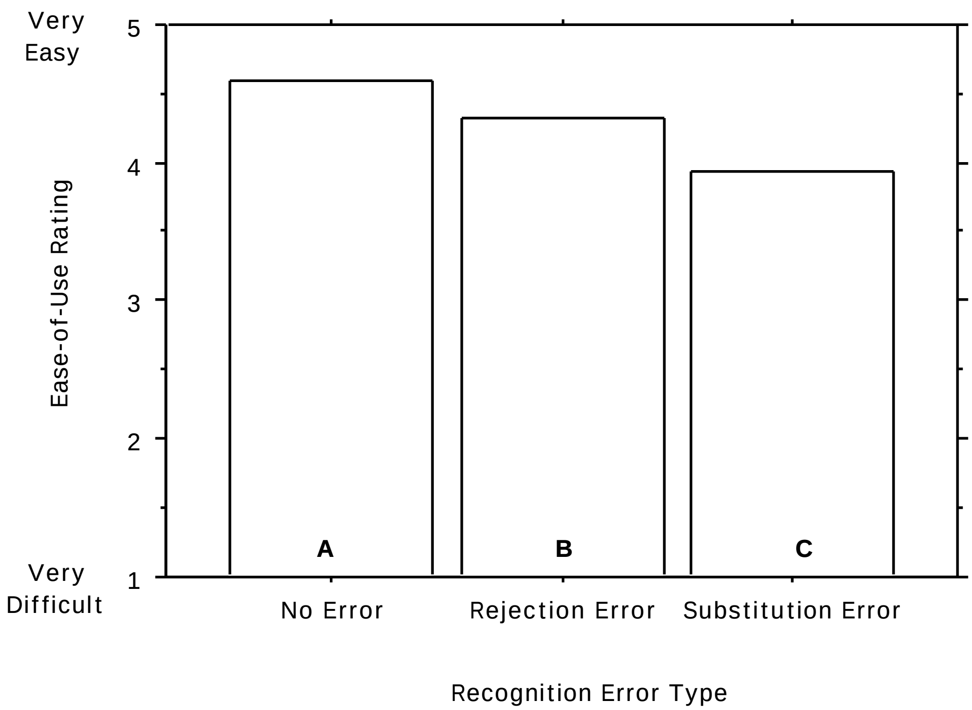

Figure 9. Mean ease-of-use rating as a function of recognition error type. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

The error correction method by error type interaction shows that the "Simply Repeat" error correction method condition resulted in higher ease-of-use ratings than the "Cancel and Repeat" error correction method when substitution errors occurred (see Figure 10). Drivers did not have to employ an error correction method under the no error condition. Under the rejection error condition, drivers were instructed through system feedback to repeat the command input. Therefore, both the no error and rejection error conditions are not relevant in this interaction. A simple effects test found an effect of error correction method on ease-of-use rating ($\underline{F}(1,9) = 4.29$, $\underline{p} < 0.100$) when substitution errors occurred. Drivers preferred repeating command inputs to correct substitution errors over canceling an error and then repeating the desired command inputs.

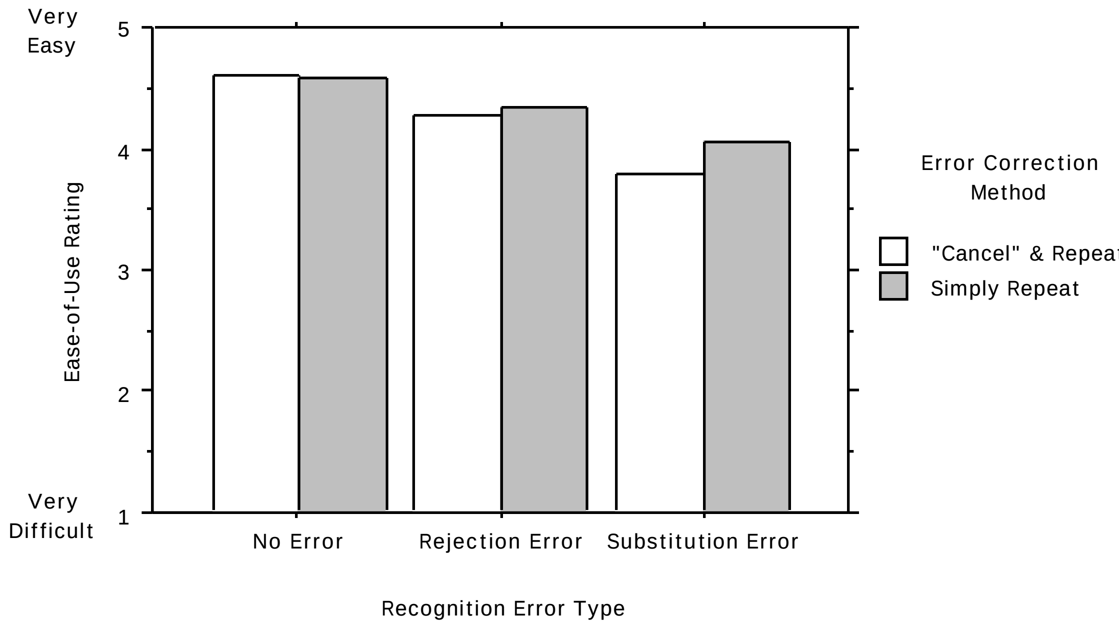

Figure 10. Mean ease-of-use rating as a function of recognition error type and error correction method.

The second question a driver was asked after completing a task was, "Using the five-point comfort scale, how comfortable did you feel performing this task while driving the vehicle?" The results from the "comfort" question were analyzed in the full-factorial MANOVA against the independent variables feedback modality, error correction method, and error type. The main effect of error type was found to be significant ($\underline{F}(2,18) = 13.75$, $\underline{p} < 0.001$). A complete ANOVA table for comfort rating is shown in Appendix H, Table H-12.

The significant error type effect is shown in Figure 11. A Student Newman-Keuls post-hoc test for error type found statistical differences between all three error type conditions. The no error condition resulted in the highest comfort ratings, followed by the rejection error condition, and then by the substitution error condition.

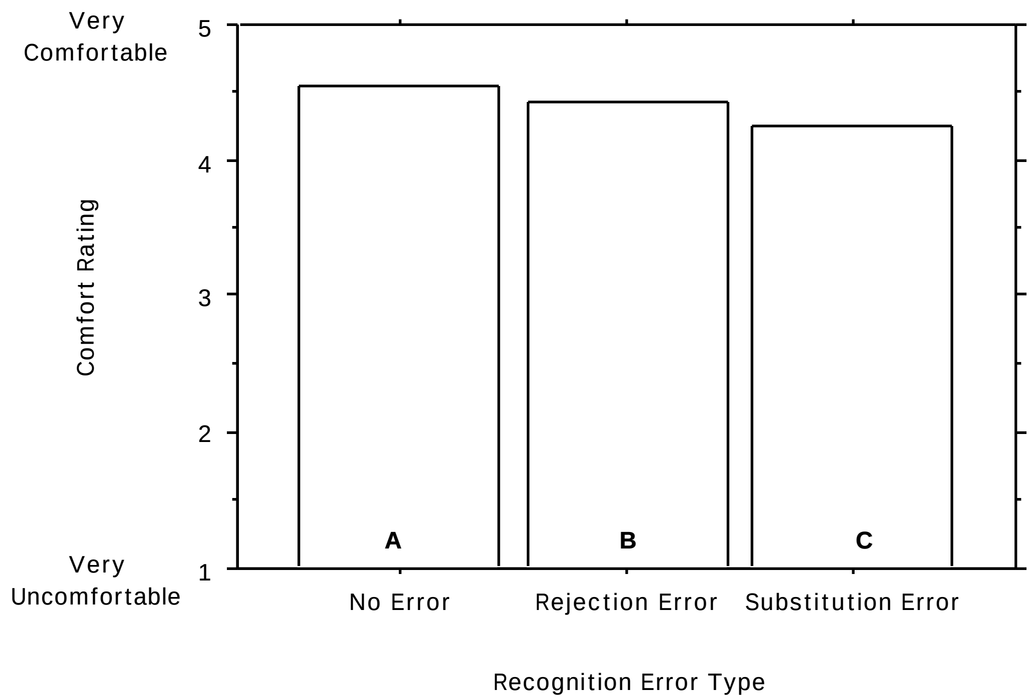

Figure 11. Mean comfort rating as a function of recognition error type. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

The final question a driver was asked after completing a task was, "Using the five-point distraction scale, how distracting was performing this task while driving the vehicle?" The results from the "distraction" question were analyzed in the full-factorial MANOVA against the independent variables feedback modality, error correction method, and error type. The main effect, error type, was significant ($\underline{F}(2,18) = 43.88$, $\underline{p} < 0.001$), as were the feedback modality by error correction method interaction ($\underline{F}(2,17) = 3.08$, $\underline{p} < 0.100$), and the feedback modality by recognition error type interaction ($\underline{F}(4,36) = 2.13$, $\underline{p} < 0.100$). A complete ANOVA table for distraction rating is shown in Appendix H, Table H-13.

The significant error type effect, feedback modality by error correction method interaction, and feedback modality by recognition error type interaction are shown in Figures 12, 13, and 14, respectively. A Student Newman-Keuls post-hoc test for the error type main effect found statistical differences between all three error type conditions. The no error condition resulted in the lowest levels of reported distraction, followed by the rejection error condition, and then by the substitution error condition. It is important to note that larger the distraction ratings indicate the task was less distracting to perform (see Appendix F for scale example).

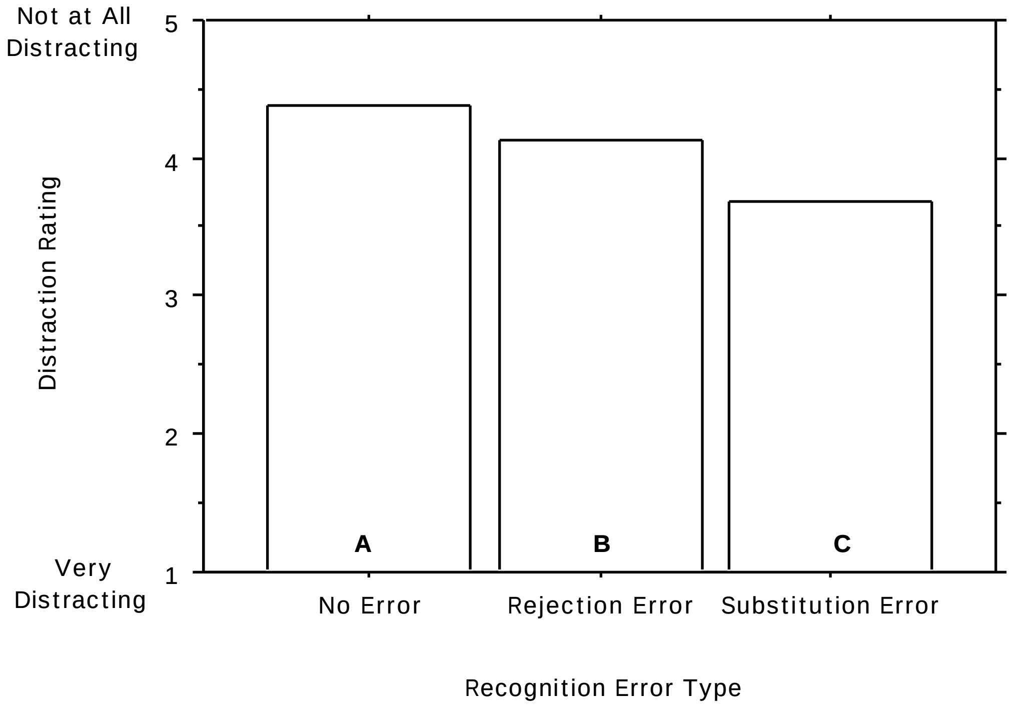

Figure 12. Mean distraction rating as a function of recognition error type. Higher rating numbers indicate less distraction. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

The feedback modality by error correction method interaction showed that the "Simply Repeat" error correction method condition results in lower levels of reported distraction for the auditory-only feedback modality and the visual-only feedback modality (see Figure 13). However, no difference in reported distraction between error correction methods was found under the visual plus auditory feedback modality.

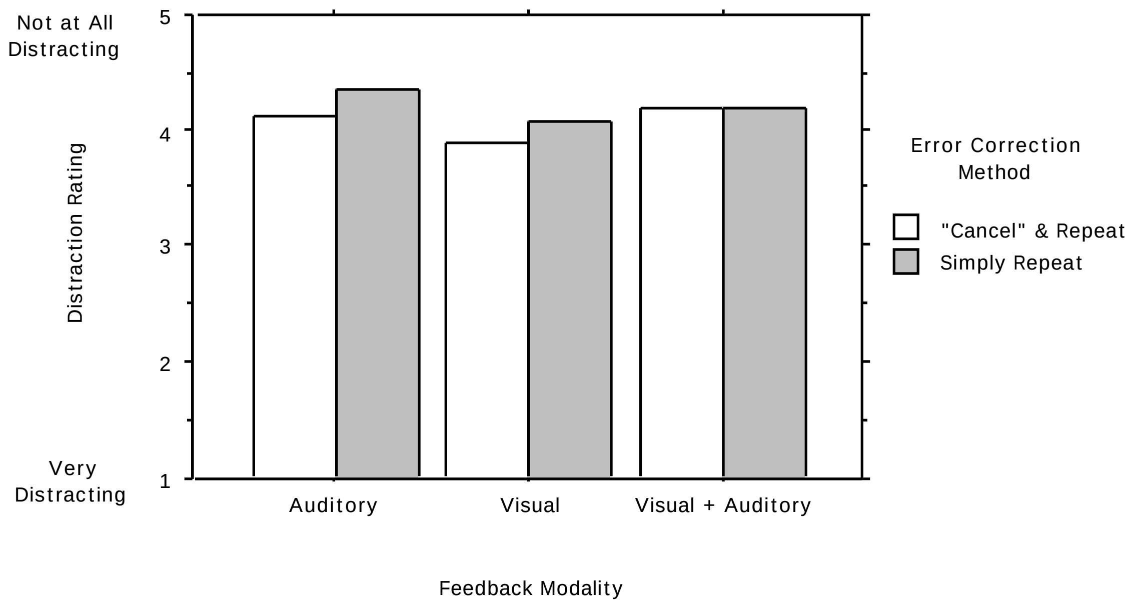

Figure 13. Mean distraction rating as a function of feedback modality and error correction method. Higher rating numbers indicate less distraction.

The feedback modality by recognition error type interaction showed the visual only feedback resulted in the highest levels of reported distraction across all recognition error types (see Figure 14). The auditory only feedback condition resulted in the lowest levels of reported distraction when either no error occurred or when a substitution error occurred while performing in-vehicle tasks. However, visual plus auditory feedback resulted in the lowest levels of reported distraction when a rejection error occurred during task performance.

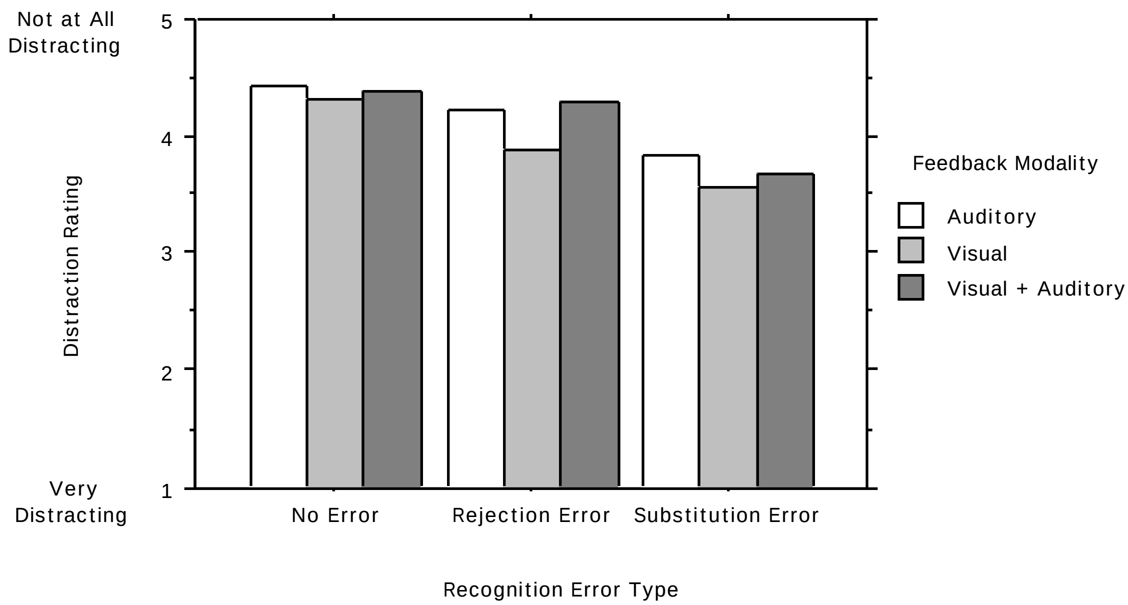

Figure 14. Mean distraction rating as a function of feedback modality and recognition error type. Higher rating numbers indicate less distraction.

Results from the measures of driver preference/acceptance showed an inter-correlation between the subjective constructs. This suggests that all three dependent measures (ease-of-use, comfort, and distraction) were assessing the same subjective construct. There may have been only a single construct that was evaluated by subjects when performing an in-vehicle task. Future research could employ only one of the of the subjective measures since it appears that ease-of-use, comfort, and distraction are all correlated. The distraction question is offered as the choice for evaluating measures of driver preference/acceptance.

Experiment 2

Analyses were performed using two sets of independent variables. The first set of independent variables analyzed was age, input condition, and in-vehicle task type. As described in the Experimental Methodology section, there were two levels of age tested: younger drivers and older drivers. The input condition variable had eight levels which were comprised of seven speech-input conditions (six levels are the two error types by three accuracy levels full-factorial, plus the 100% recognition accuracy control condition) and one manual-input condition. Sixteen different in-vehicle task types were performed by subjects for each combination of the independent variables. Effects of age and input condition on driving performance, task-function usability, and driver preference/acceptance were determined from a mixed-factorial multivariate analysis of variance. Significant MANOVA results were analyzed further in separate univariate analyses of variance. All significant task type by input condition interactions were analyzed using simple effects tests to evaluate the reliability of the decision trees in determining the overall input-modality allocation recommendation.

The second set of independent variables analyzed was age, recognition accuracy, and error type. Data from the manual-input and 100% recognition accuracy control conditions were dropped from this analysis so that a mixed-factorial MANOVA could be performed on the remaining independent variables to determine if any second-order interactions exist between age, recognition accuracy, and error type. Second-order interactions may prove helpful in the development of human factors recommendations and design guidelines for the use of ASR technology in automobiles. Again, separate ANOVAs were performed on the dependent measures whenever significant MANOVA results were found.

Due to the number of statistical tests performed in Experiment 2, the criterion set for statistical significance was changed to $\alpha = 0.05$ to reduce the probability of spuriously significant results. Therefore, familywise Type I error rate is controlled at an acceptable level ($\alpha_{FW} = 0.10$) for the two comparisons performed on data from Experiment 2.

Analysis of variance tables are presented in Appendix I for the driving performance data, task-function usability data, and driver preference/acceptance data. Only significant main effects and two-way interactions are reported and discussed in this section. It is important to note that a number of F-ratios were less than 1.0 for the ANOVA results in Experiment 2. F-ratios less than one indicate the likelihood for heterogeneity of variance between sample populations. An assumption of the ANOVA technique is that each of the sample populations of scores has the same variance (Howell, 1992). Although analysis of variance is a very robust statistical procedure, violating the assumption of homogeneity of variance can lead to an inflation in Type I error (Keppel, 1991). The reader is cautioned of this fact; therefore, it is important not to place too great an emphasis on any single significant result because the probability of finding a difference between levels of the independent variable, when no difference actually exists, is increased.

Measures of driving performance. Three major categories of dependent measures were used in evaluating driving performance. The categories include (1) eye movement behavior, (2) steering input behavior, and (3) speed maintenance behavior. Results are presented for each category when significant main effects or interactions were found. The significance of the results and the implications they may have on ASR system design are also discussed in this section when appropriate.

Eye movement behavior refers to the location and the duration a driver is looking while driving, or driving and performing an in-vehicle task. Percent eyes-off-road time was the measure used for determining the effect of age, input condition, and task type or age, recognition accuracy, and error type on driver eye-movement behavior. Percent eyes-off-road time is an important measure for data collected in Experiment 2 because differences in driver eye-movement behavior may be shown between driving while performing manual-input tasks and driving while performing similar speech-input tasks. A claimed benefit for using ASR technology in automotive applications is that drivers should have fewer in-vehicle glances and for shorter durations when performing speech-input tasks compared with current manual-input tasks. Drivers' percent eyes-off-road time was predicted to be less with speech-input tasks than with similar manual-input tasks.

Mean percent eyes-off-road time when subjects were performing in-vehicle tasks was analyzed in two mixed-factorial MANOVAs against the independent variables (1) age, input condition, and task type; and (2) age, recognition accuracy, and error type. Two main effects were significant: input condition ($\underline{F}(7,56) = 16.34$, $\underline{p} < 0.001$) and recognition accuracy ($\underline{F}(2,20) = 6.28$, $\underline{p} < 0.010$). The task type by input condition interaction was also significant ($\underline{F}(120,1068) = 1.35$, $\underline{p} < 0.010$). Complete ANOVA tables for percent eyes-off-road time are shown in Appendix I, Table I-1 and I-2.

The significant input condition and recognition accuracy main effects are shown in Figures 15 and 16, respectively. A Student Newman-Keuls post-hoc test for the input condition effect found statistical differences between the manual-input condition and all seven speech-input conditions. No statistical difference was found between any of the speech-input conditions (see Figure 15). Performing in-vehicle tasks using speech input resulted in significantly greater mean percent eyes-off-road time than performing identical tasks with manual input. This is contrary to what was hypothesized for speech-input tasks.

A possible explanation might be that the novelty of using speech input to perform in-vehicle tasks caused drivers to look inside the vehicle for longer periods of time to confirm that an operation was correctly carried out by the ASR system. A similar novelty effect was found for head-up displays (HUDs) when subjects drove a HUD-equipped vehicle for the first time (Kiefer, 1991). Drivers had an initial tendency to look more often at the HUD than the conventional speedometer display (i.e., drivers demonstrated greater percent eyes-off-road time initially with a HUD compared with a conventional instrument panel speedometer display). However, after limited HUD experience (approximately 35 minutes), drivers showed no significant difference in percent eyes-off-road time compared with a conventional speedometer display equipped vehicle (Kiefer, 1991).

A second explanation might be a practice effect. After a longer period of practice with an ASR system, drivers might demonstrate lower percent eyes-off-road time when performing speech-input tasks. The percent eyes-off-road time for speech-input tasks with adequate practice may approximate, or become less than, the percent eyes-off-road for manual-input tasks. The manual-input tasks performed in this experiment may have been similar to manual tasks that the drivers commonly performed in their own vehicles. Drivers would have had more practice performing manual-input tasks than performing speech-input tasks, prior to participating in the experiment. Increased practice with the manual-input tasks may have allowed drivers to maintain lower percent eyes-off-road times.

A final explanation for why mean percent eyes-off-road time was greater for speech-input tasks may have something to do with the type of feedback modality provided by the ASR system. A visual plus auditory feedback method was used in Experiment 2. The visual component of the feedback may have compelled drivers to look inside the vehicle during task performance, resulting in the greater percent eyes-off-road time for speech-input tasks recorded. Further research would be required to determine if an auditory-only type of feedback could reduce eyes-off-road time compared to a visual plus auditory feedback method.

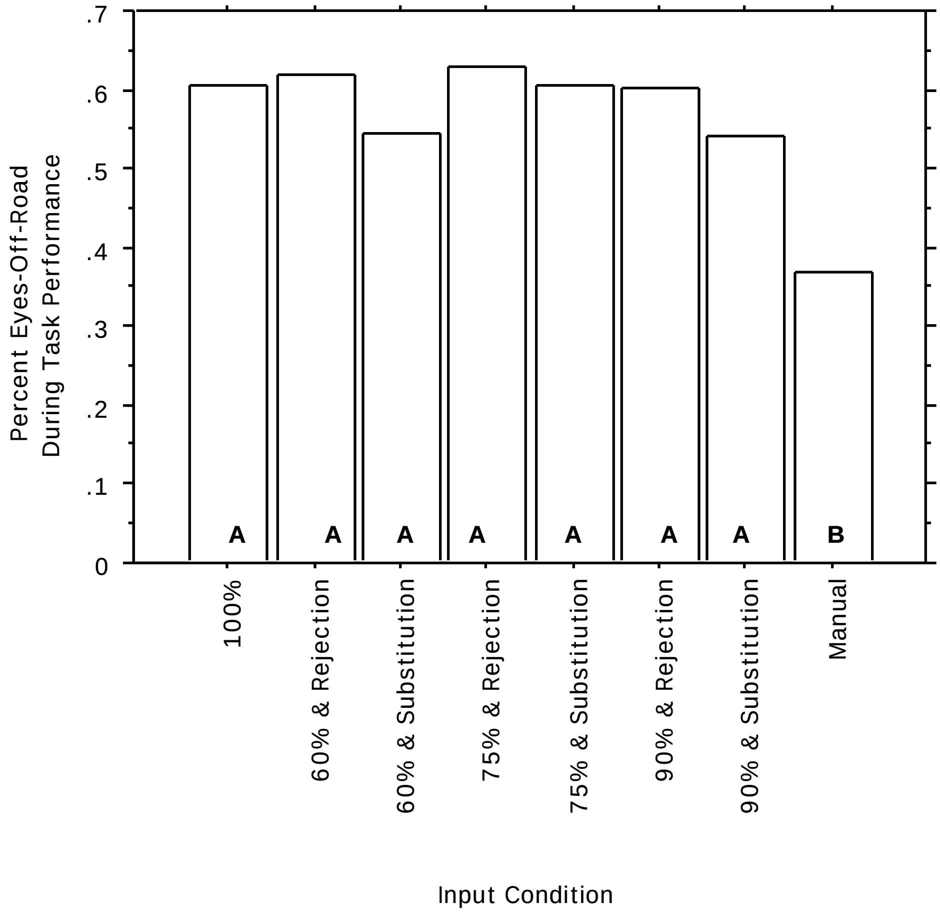

Figure 15. Mean percent eyes-off-road during task performance as a function of input condition. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

A Student Newman-Keuls post-hoc test for the recognition accuracy effect showed that both the 90% and 60% recognition accuracy conditions result in statistically smaller percent eyes-off-road time than the 75% recognition accuracy condition. No statistical difference was found between the 90% and 60% recognition accuracy conditions (see Figure 16).

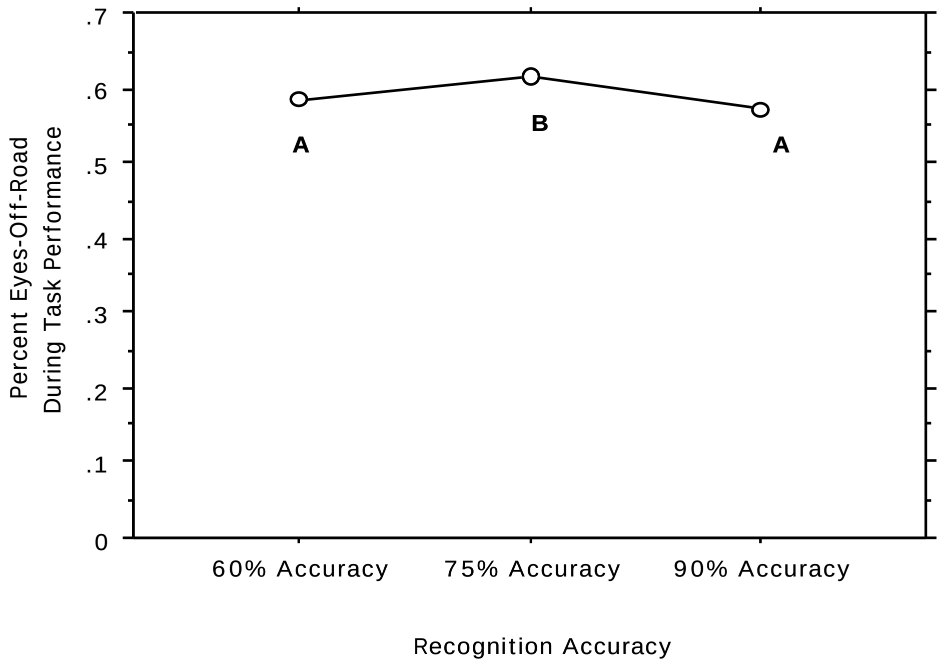

Figure 16. Mean percent eyes-off-road during task performance as a function of recognition accuracy. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

Results from the simple effects tests for the task type by input condition interaction are shown in Table 4. In general, the results contradict the overall decision tree recommendations for only two of the task types: (1) change climate mode (Task Type number 8); and (2) tune radio to preset station (Task Type number 11). The results showed that the "climate mode" task and the "radio preset" task should be performed using manual input instead of the recommended speech input to minimize drivers' percent eyes-off-road. A possible explanation for why the manual condition resulted in smaller percent eyes-off-road time compared with the speech conditions for these two tasks is that it was difficult for drivers to confirm if the task was performed correctly using only the feedback from the task-function's display. The display indicating the radio preset number was much smaller in font size than the radio's frequency numbers. Using speech input, the drivers only feedback that the radio was changed to the specified preset station was this display. For the climate mode task, drivers often reported forgetting exactly where they needed to look to see if the task was performed correctly. The climate system had two round knobs and one round display that all looked similar. Small display fonts and confusion between similar-looking displays could lead to increased percent eyes-off-road when drivers tried to confirm that a task was performed correctly. The decision trees did not consider such design characteristics in their development. The remaining task types had statistically similar results between the manual-input condition and one or more of the speech-input conditions. Design tradeoffs for whether the remaining tasks should be performed using speech or manual input could be determined by other criteria such as cost or ease of implementation in place of the decision tree recommendations.

Table 4. Results from the simple effects tests for the significant task type by input condition interaction for the percent eyes-off-road dependent measure. (Task Type number corresponds with the numbers for in-vehicle tasks listed described in Table 3 of the Experimental Methodology Section.)

Task Type	**Overall Decision Tree Recommendation**	**Manual**	**100%**	**90% w/ Rej. Error**	**90% w/ Sub. Error**	**75% w/ Rej. Error**	**75% w/ Sub. Error**	**60% w/ Rej. Error**	**60% w/ Sub. Error**
1	Manual	A	A	A	A	A	A	A	A
means		0.33	0.49	0.65	0.31	0.63	0.48	0.65	0.39
2	Speech	A	A	A	A	A	A	A	A
means		0.38	0.63	0.45	0.60	0.63	0.62	0.69	0.53
3	Manual	A	A	A	A	A	A	A	A
means		0.44	0.67	0.65	0.44	0.70	0.62	0.59	0.63
4	Manual	A	A	A	A	A	A	A	A
means		0.46	0.60	0.55	0.63	0.61	0.62	0.56	0.53
5	Manual	A	A	A	A	A	A	A	A
means		0.42	0.53	0.45	0.62	0.44	0.73	0.49	0.66
6	Manual	C	AB	AB	AB	A	AB	A	B
means		0.27	0.61	0.60	0.60	0.79	0.60	0.78	0.53
7	Speech	C	AB	A	BC	A	ABC	ABC	ABC
means		0.37	0.59	0.67	0.42	0.68	0.55	0.56	0.51
8	Speech	B	A	A	A	A	A	A	A
means		0.19	0.57	0.62	0.61	0.55	0.65	0.64	0.57
9	Speech	B	AB	A	AB	A	A	A	AB
means		0.37	0.50	0.58	0.49	0.60	0.59	0.67	0.46
10	Speech	B	AB	AB	AB	A	AB	AB	AB
means		0.40	0.58	0.58	0.48	0.70	0.53	0.64	0.50
11	Speech	B	A	A	A	A	A	A	A
means		0.29	0.81	0.69	0.69	0.73	0.54	0.58	0.60
12	Manual	B	AB	A	AB	AB	AB	A	AB
means		0.31	0.61	0.78	0.52	0.59	0.63	0.75	0.57
13	Speech	B	AB	A	AB	AB	AB	A	AB
means		0.32	0.56	0.67	0.62	0.62	0.56	0.77	0.54
14	Manual	A	A	A	A	A	A	A	A
means		0.50	0.62	0.52	0.41	0.45	0.66	0.49	0.49
15	Speech	A	A	A	A	A	A	A	A
means		0.46	0.67	0.60	0.62	0.66	0.62	0.40	0.64
16	Speech	A	A	A	A	A	A	A	A
means		0.40	0.57	0.65	0.60	0.65	0.60	0.63	0.51

Steering input behavior was evaluated through measurements of steering variability and the number of steering wheel reversals greater than a specified number of degrees of steering wheel rotation (e.g., 6 and 12 degrees). Steering variability (i.e., variance), while subjects performed in-vehicle tasks, was analyzed in the mixed-factorial MANOVA against the independent variables age, input condition, and task type. Only one main effect was significant: age, $F(1,10) = 53.35$, $p < 0.001$. A similar effect for age was significant ($F(1,10) = 22.28$, $p < 0.001$) when steering variance was analyzed in the mixed-factorial MANOVA against the independent variables age, recognition accuracy, and error type. Complete ANOVA tables for steering variability are shown in Appendix I, Table I-3 and I-4. The effect of age on steering variability shows that steering variance is significantly greater for younger drivers when performing in-vehicle tasks than for older drivers (steering variance for younger drivers = 0.745 degrees; steering variance for older drivers = 0.510 degrees).

The number of steering-wheel-reversals per minute greater than 6 degrees, and the number of steering-wheel-reversals per minute greater than 12 degrees were analyzed in the mixed-factorial MANOVA against the independent variables age, input condition, and task type. The main effects, age ($F(1,10) = 5.85$, $p < 0.050$) and input condition ($F(7,70) = 3.82$, $p < 0.005$), were found to be significant for the "Steering Reversals > 6°" analysis, as was the task by input condition interaction ($F(120,1292) = 2.72$, $p < 0.001$). The "Steering Reversals > 12°" analysis found similar effects of age ($F(1,10) = 9.28$, $p < 0.025$) and input condition ($F(7,70) = 2.92$, $p < 0.010$), including a significant task by input condition interaction ($F(120,1292) = 1.53$, $p < 0.001$).

The number of steering reversals per minute greater than 6 and 12 degrees were also analyzed in the mixed-factorial MANOVA performed on the age, recognition accuracy, and error type independent variables. The "Steering Reversals > 6°" analysis resulted in significant effects for both age ($F(1,10) = 7.25$, $p < 0.025$) and recognition accuracy ($F(2,20) = 13.24$, $p < 0.001$). The "Steering Reversals > 12°" analysis resulted in similar effects for age ($F(1,10) = 7.72$, $p < 0.025$) and recognition accuracy ($F(2,20) = 11.09$, $p < 0.001$). Complete ANOVA tables for the steering reversals greater than 6 and 12 degrees dependent measures are presented in Appendix I, Tables I-5, I-6, I-7, and I-8.

The effect of age on steering reversals per minute greater than 6 degrees showed that younger drivers had significantly greater numbers of steering reversals per minute when performing in-vehicle tasks than older drivers (number of steering reversals > 6° for younger drivers = 4.70 per minute; number of steering reversals > 6° for older drivers = 3.55 per minute). The effect of age on steering reversals per minute greater than 12 degrees also showed that younger drivers had significantly greater numbers of steering reversals per minute when performing in-vehicle tasks than older drivers (number of steering reversals > 12° for younger drivers = 1.29 per minute; number of steering reversals > 12° for older drivers = 0.93 per minute).

The significant input condition and recognition accuracy main effects for the "Steering Reversals > 6°" analysis are shown in Figures 17 and 18, respectively. A Student Newman-Keuls post-hoc test for the input condition effect found statistical differences between the 90% accuracy with rejection errors condition and both the manual-input condition and the 60% accuracy with substitution errors condition. The manual-input condition was statistically similar to all other speech input conditions. The 90% accuracy with rejection errors condition resulted in the fewest steering reversals, while the 60% accuracy with substitution errors condition resulted in the greatest number of steering reversals (see Figure 17).

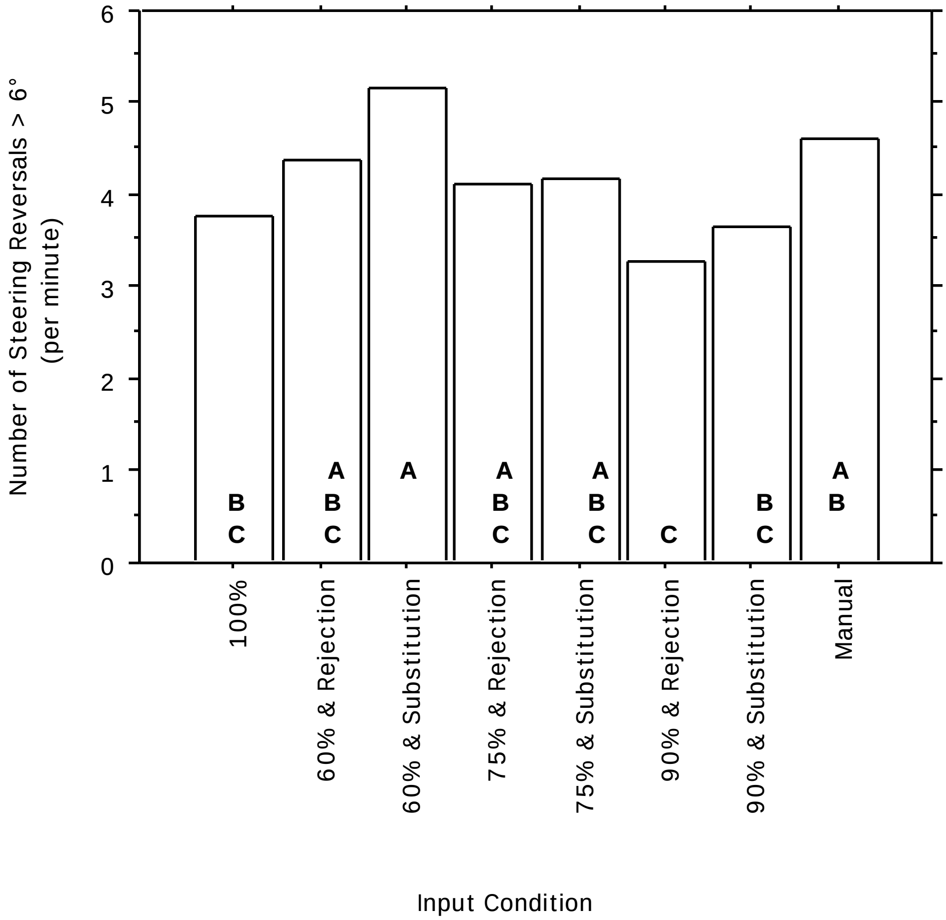

Figure 17. Number of steering reversals per minute greater than 6 degrees during task performance as a function of input condition. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

A Student Newman-Keuls post-hoc test for the recognition accuracy effect showed that all three accuracy conditions were statistically different from one another. The 90% recognition accuracy condition resulted in the fewest number of steering reversals, while the 60% accuracy condition resulted in the greatest number of reversals greater than 6 degrees (see Figure 18).

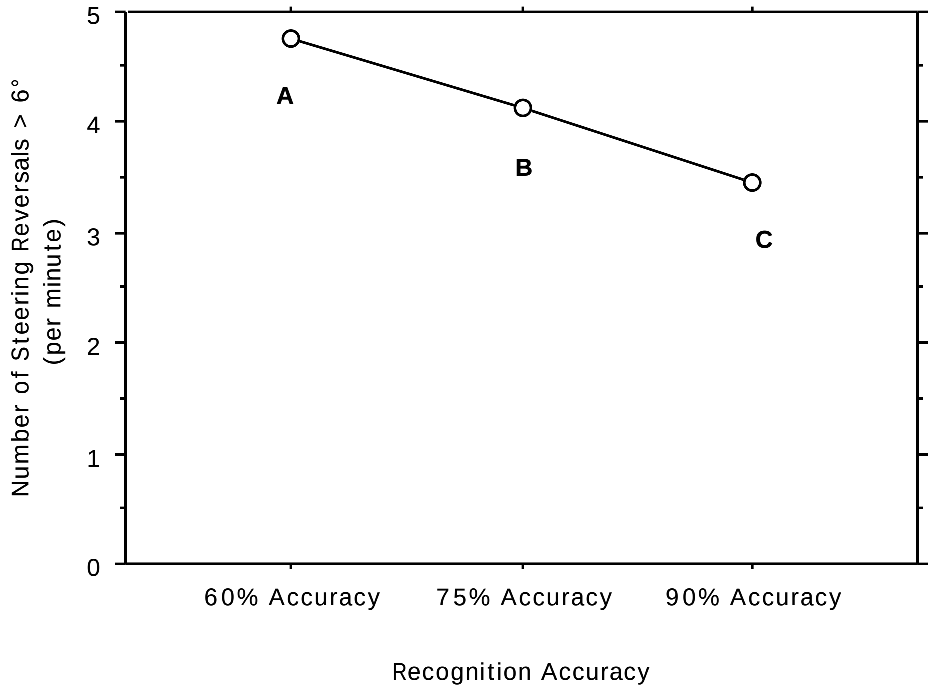

Figure 18. Number of steering reversals per minute greater than 6 degrees during task performance as a function of recognition accuracy. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

Results from the simple effects tests for the task type by input condition interaction are shown in Table 5 for the "Steering Reversals > 6°" analysis. The results showed that all task types had statistically similar results between the manual-input condition and one or more of the speech-input conditions. Design tradeoffs for whether the tasks should be performed using speech or manual input could be determined by other criteria such as cost or ease of implementation in place of the decision tree recommendations.

Table 5. Results from the simple effects test for the significant task type by input condition interaction for the number of steering reversals per minute greater than 6 degrees dependent measure. (Task Type number corresponds with the numbers for in-vehicle tasks listed described in Table 3 of the Experimental Methodology Section.)

Task Type	**Overall Decision Tree Recommendation**	**Manual**	**100%**	**90% w/ Rej. Error**	**90% w/ Sub. Error**	**75% w/ Rej. Error**	**75% w/ Sub. Error**	**60% w/ Rej. Error**	**60% w/ Sub. Error**
1	Manual	A	A	A	A	A	A	A	A
means		2.74	2.57	1.96	3.62	3.82	2.33	2.83	4.30
2	Speech	A	A	A	A	A	A	A	A
means		4.37	4.25	4.37	3.67	4.29	3.36	4.62	4.81
3	Manual	B	B	B	A	B	B	B	B
means		2.20	1.99	1.58	4.02	2.50	2.60	2.30	2.05
4	Manual	A	A	A	A	A	A	A	A
means		5.08	4.38	5.14	4.56	6.52	5.09	6.93	12.56
5	Manual	B	B	B	B	A	B	B	B
means		2.80	2.4	1.91	2.25	4.62	2.54	2.21	1.99
6	Manual	A	A	A	A	A	A	A	A
means		5.52	4.84	3.34	3.45	3.63	5.42	3.68	5.49
7	Speech	A	A	A	A	A	A	A	A
means		3.47	3.71	3.96	3.33	4.71	6.28	5.05	5.18
8	Speech	A	A	A	A	A	A	A	A
means		5.06	2.49	3.27	4.51	3.47	4.58	4.37	6.15
9	Speech	A	A	A	A	A	A	A	A
means		6.83	7.54	2.86	2.91	3.46	4.69	4.11	6.01
10	Speech	A	A	A	A	A	A	A	A
means		4.23	2.75	2.39	3.19	2.86	4.14	3.45	4.78
11	Speech	A	A	A	A	A	A	A	A
means		4.89	3.43	2.61	3.78	3.81	3.28	5.17	3.84
12	Manual	AB	B	AB	AB	AB	AB	AB	A
means		5.43	2.82	4.09	4.48	3.81	3.63	5.48	6.43
13	Speech	AB	AB	AB	B	AB	AB	A	AB
means		3.59	2.82	3.10	2.21	3.11	3.60	4.84	4.19
14	Manual	A	A	A	A	A	A	A	A
means		2.72	3.34	2.65	4.12	3.28	4.12	3.28	3.22
15	Speech	A	A	A	A	A	A	A	A
means		7.15	5.06	4.22	4.06	5.59	5.87	5.39	8.89
16	Speech	A	A	A	A	A	A	A	A
means		7.49	6.91	4.62	4.44	6.12	5.50	6.45	3.95

The significant input condition and recognition accuracy main effects for the "Steering Reversals > 12°" analysis are shown in Figures 19 and 20, respectively. A Student Newman-Keuls post-hoc test for the input condition effect found a statistical difference between the 60% accuracy with substitution errors condition and the remaining input conditions. The remaining input conditions were all statistically similar to each other. The 60% accuracy with substitution errors condition resulted in the greatest number of steering reversals (see Figure 19).

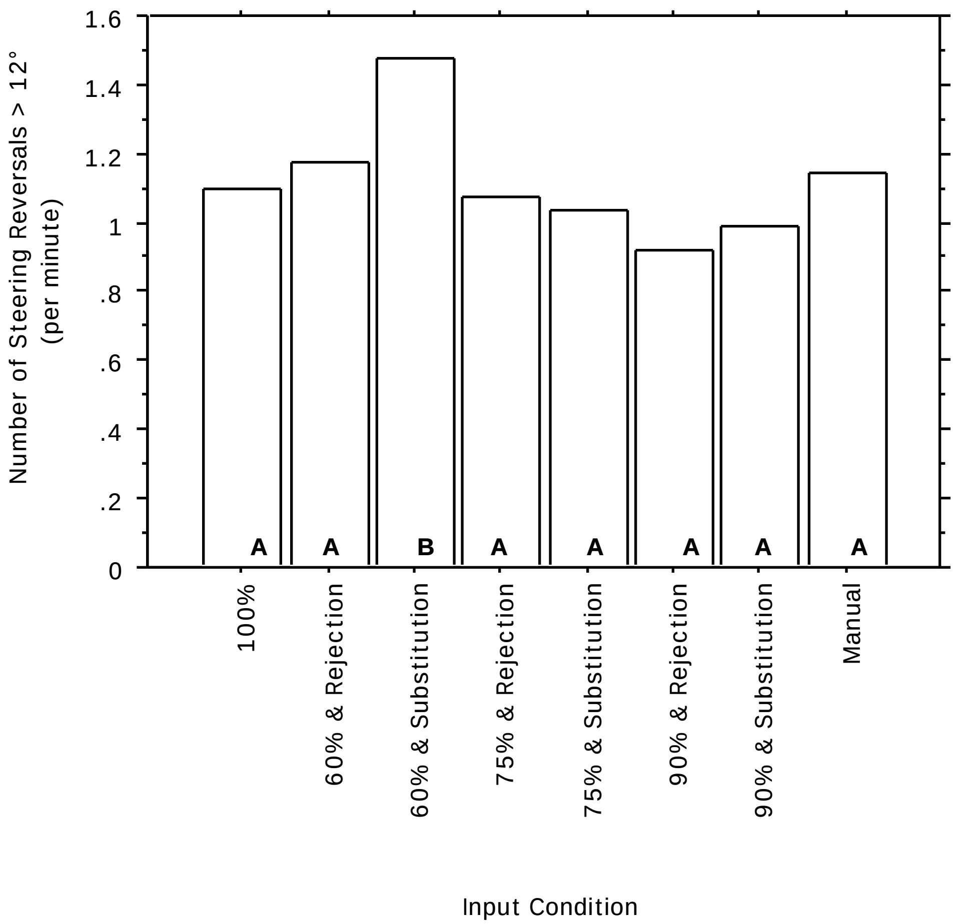

Figure 19. Number of steering reversals per minute greater than 12 degrees during task performance as a function of input condition. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

A Student Newman-Keuls post-hoc test for the recognition accuracy effect showed that the 60% recognition accuracy condition resulted in a significantly larger number of steering reversals greater than 12 degrees compared with both the 75% and 90% recognition accuracy conditions. The 90% and 75% recognition accuracy conditions were statistically similar (see Figure 20).

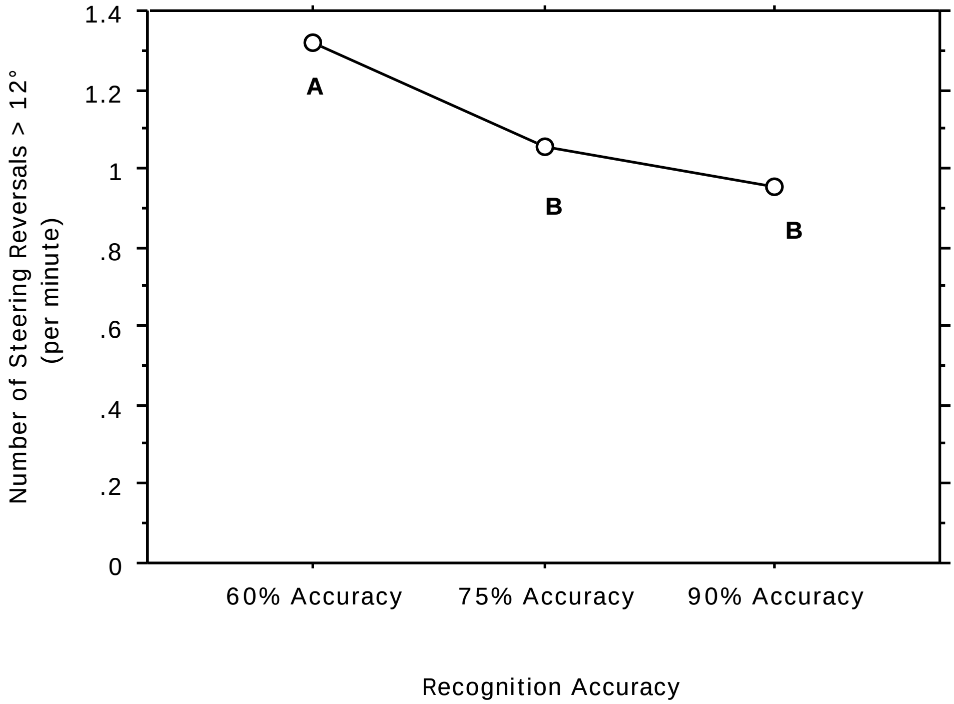

Figure 20. Number of steering reversals per minute greater than 12 degrees during task performance as a function of recognition accuracy. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

Results from the simple effects tests for the task type by input condition interaction are shown in Table 6 for the "Steering Reversals > 12°" analysis. The results showed that all task types had statistically similar results between the manual-input condition and all of the speech-input conditions. Design tradeoffs for whether the tasks should be performed using speech or manual input could be determined by other criteria such as cost or ease of implementation in place of the decision tree recommendations.

Table 6. Results from the simple effects test for the significant task type by input condition interaction for the number of steering reversals per minute greater than 12 degrees dependent measure. (Task Type number corresponds with the numbers for in-vehicle tasks listed described in Table 3 of the Experimental Methodology Section.)

Task Type	**Overall Decision Tree Recommendation**	**Manual**	**100%**	**90% w/ Rej. Error**	**90% w/ Sub. Error**	**75% w/ Rej. Error**	**75% w/ Sub. Error**	**60% w/ Rej. Error**	**60% w/ Sub. Error**
1	Manual	A	A	A	A	A	A	A	A
means		1.01	0.68	0.38	1.07	1.31	0.75	0.95	1.22
2	Speech	A	A	A	A	A	A	A	A
means		1.09	1.45	1.60	0.77	1.05	1.19	0.61	1.78
3	Manual	A	A	A	A	A	A	A	A
means		0.62	0.73	0.25	0.73	0.68	0.87	0.79	0.37
4	Manual	A	A	A	A	A	A	A	A
means		1.50	1.18	1.61	1.38	2.45	1.42	2.07	3.79
5	Manual	A	A	A	A	A	A	A	A
means		1.18	0.90	0.47	0.76	1.18	0.73	0.91	0.47
6	Manual	A	A	A	A	A	A	A	A
means		1.20	1.27	0.97	0.76	0.77	1.54	0.82	1.84
7	Speech	A	A	A	A	A	A	A	A
means		1.05	0.76	0.62	0.75	1.53	1.38	1.16	0.80
8	Speech	A	A	A	A	A	A	A	A
means		1.35	0.95	0.82	1.21	1.08	1.18	1.26	1.62
9	Speech	A	A	A	A	A	A	A	A
means		1.51	1.30	1.15	0.59	0.65	1.28	0.97	1.85
10	Speech	A	A	A	A	A	A	A	A
means		0.91	0.63	0.63	0.86	0.34	1.05	1.40	1.27
11	Speech	A	A	A	A	A	A	A	A
means		1.26	1.49	0.49	0.80	0.73	1.02	1.75	1.01
12	Manual	A	A	A	A	A	A	A	A
means		1.28	1.05	1.49	2.19	0.99	0.98	1.22	2.12
13	Speech	A	A	A	A	A	A	A	A
means		1.09	0.96	0.69	0.75	1.02	0.66	1.31	1.32
14	Manual	A	A	A	A	A	A	A	A
means		0.86	1.02	1.35	1.60	0.87	0.73	0.75	1.18
15	Speech	A	A	A	A	A	A	A	A
means		1.28	1.94	0.67	0.80	1.09	0.75	1.22	2.31
16	Speech	A	A	A	A	A	A	A	A
means		1.07	1.29	1.42	0.68	1.41	0.98	1.59	0.62

Speed maintenance behavior was evaluated through measures of accelerator pedal position variability, mean vehicle velocity (in miles per hour), vehicle velocity variability (in miles per hour), and the number of brake activations per minute. All measures of speed maintenance behavior analyzed were recorded while drivers were performing in-vehicle tasks.

Accelerator position variance was analyzed in the two mixed-factorial MANOVAs against the independent variables (1) age, input condition, and task type; and (2) age, recognition accuracy, and error type. Only one main effect was significant: age, (1) $\underline{F}(1,10) = 37.73$, $\underline{p} < 0.001$, and (2) $\underline{F}(1,10) = 34.46$, $\underline{p} < 0.001$, respectively. Complete ANOVA tables for accelerator position variability are shown in Appendix I, Tables I-9 and I-10.

The effect of age on accelerator position variance showed that accelerator position variance was significantly greater for younger drivers when performing in-vehicle tasks than for older drivers (accelerator position variance--normalized between 0 and 1--for younger drivers = 0.0170 ; accelerator position variance--normalized between 0 and 1--for older drivers = 0.0129).

Mean vehicle velocity and vehicle velocity variance during task performance were analyzed in the two mixed-factorial MANOVAs against the independent variables (1) age, input condition, and task type; and (2) age, recognition accuracy, and error type. No significant main effects or interactions were for either dependent measure. Complete ANOVA tables are presented in Appendix I, Tables I-11, I-12, I-13, and I-14.

Finally, the number of brake activations per minute was analyzed in the two full-factorial MANOVAs against the independent variables (1) age, input condition, and task type; and (2) age, recognition accuracy, and error type. Analysis of brake activations provided further information about driving performance when subjects were driving and concurrently performing in-vehicle tasks under different input modalities and ASR system parameters. Only one significant main effect was found from the analyses of brake activations: age, (1) $\underline{F}(1,10) = 35.74$, $\underline{p} < 0.001$, and (2) $\underline{F}(1,10) = 66.09$, $\underline{p} < 0.001$, respectively. A significant age by input condition interaction ($\underline{F}(7,70) = 2.28$, $\underline{p} < 0.050$) and a significant task type by input condition interaction ($\underline{F}(120,1292) = 3.07$, $\underline{p} < 0.001$) were found from the analyses. Complete ANOVA tables for brake activations are shown in Appendix I, Table I-15 and I-16.

The significant age effect showed that older drivers demonstrated larger numbers of brake activations per minute than younger drivers (number of brake activations per minute for younger drivers = 1.10; number of brake activations per minute for older drivers = 2.43). Brake activations may have been due in part to the road geometry on which tasks were performed. U.S. Route 460 west between Blacksburg, VA and Princeton, WV has numerous steep grades that must be driven over as the road winds through the Appalachian Mountains. Brake activations were necessary at times to maintain safe vehicle speeds while traveling down the steep grades.

The significant age by input condition interaction is shown in Figure 21. The smallest differences in number of brake activations between younger drivers and older drivers were found under the manual-input and 100% recognition accuracy conditions. Differences in number of brake activations between younger and older drivers were much larger for the remaining speech input conditions.

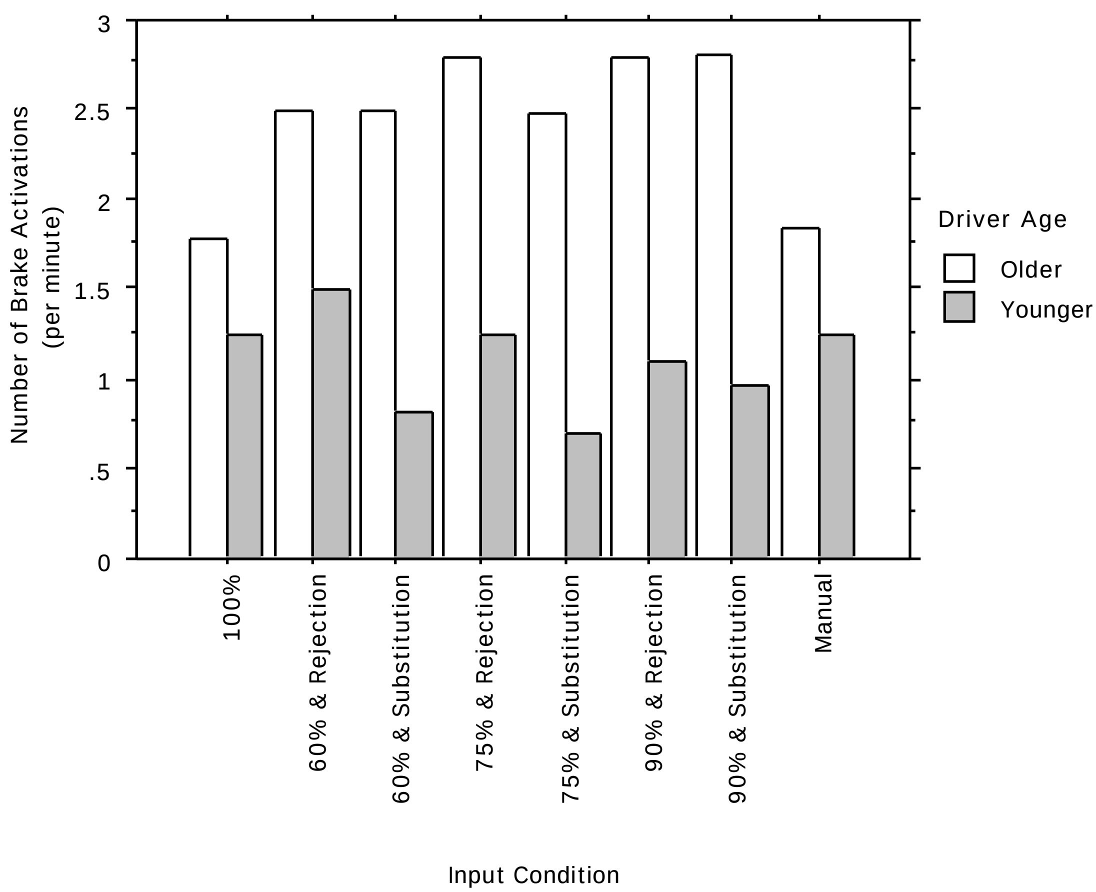

Figure 21. Number of brake activations per minute during task performance as a function of age and input condition.

Results from the simple effects tests for the task type by input condition interaction are shown in Table 7. The results showed that all task types had statistically similar results between the manual-input condition and one or more of the speech-input conditions. Design tradeoffs for whether the tasks should be performed using speech or manual input could be determined by other criteria such as cost or ease of implementation in place of the decision tree recommendations.

Table 7. Results from the simple effects test for the significant task type by input condition interaction for the number of brake activations per minute dependent measure. (Task Type number corresponds with the numbers for in-vehicle tasks listed described in Table 3 of the Experimental Methodology Section.)

Task Type	**Overall Decision Tree Recommendation**	**Manual**	**100%**	**90% w/ Rej. Error**	**90% w/ Sub. Error**	**75% w/ Rej. Error**	**75% w/ Sub. Error**	**60% w/ Rej. Error**	**60% w/ Sub. Error**
1	Manual	A	A	A	A	A	A	A	A
means		2.04	1.67	1.71	2.25	2.33	2.12	2.18	2.24
2	Speech	A	A	A	A	A	A	A	A
means		1.33	1.31	2.10	2.16	2.11	1.46	1.89	1.91
3	Manual	A	A	A	A	A	A	A	A
means		1.23	1.01	1.08	1.64	1.17	0.88	1.39	1.20
4	Manual	AB	B	AB	AB	A	AB	A	B
means		1.44	0.93	1.36	1.52	2.23	1.41	2.34	.937
5	Manual	B	AB	AB	AB	AB	B	A	B
means		1.38	1.55	1.59	1.90	1.84	1.42	2.38	1.47
6	Manual	A	A	A	A	A	A	A	A
means		1.30	1.72	2.10	1.67	1.52	1.04	1.26	1.06
7	Speech	AB	B	A	B	AB	B	AB	B
means		1.84	1.16	2.31	0.90	1.55	0.87	1.99	0.97
8	Speech	A	A	A	A	A	A	A	A
means		1.75	1.28	2.60	2.85	2.02	1.89	3.16	2.81
9	Speech	A	A	A	A	A	A	A	A
means		1.46	1.67	1.90	1.75	1.43	1.21	1.26	0.93
10	Speech	A	A	A	A	A	A	A	A
means		1.28	1.06	1.66	0.76	2.00	2.02	1.35	1.78
11	Speech	AB	AB	AB	AB	A	B	AB	AB
means		1.78	1.97	2.17	1.65	2.30	1.03	2.21	2.00
12	Manual	A	A	A	A	A	A	A	A
means		1.62	1.79	2.54	2.76	2.08	2.09	2.36	2.04
13	Speech	A	A	A	A	A	A	A	A
means		2.28	2.10	3.07	2.94	2.80	3.05	1.96	2.10
14	Manual	A	A	A	A	A	A	A	A
means		1.50	1.78	2.46	2.80	3.01	1.63	2.25	1.47
15	Speech	A	A	A	A	A	A	A	A
means		1.54	1.99	1.25	1.29	2.04	1.45	2.10	1.32
16	Speech	A	A	A	A	A	A	A	A
means		0.92	0.85	1.06	1.02	1.78	1.08	2.21	0.58

In general, the results from the measures of driving performance are indicative of younger drivers driving more aggressively than older drivers. These results support the findings from numerous other empirical studies that have shown that older drivers differ significantly from younger drivers in measures of driving performance (e.g., see Allen et al., 1991; Dingus et al., 1989, 1991,1994, and 1997; Kiefer and Gellatly, 1996; Kurokawa, 1990; Marin-Lamellet and Dejeammes, 1995; Mollenhauer et al., 1994; Paelke, 1993; Perez and Mast, 1992; Serafin et al., 1993; Sivak et al., 1991; Snyder and Monty, 1986; Walker et al., 1991; Wierwille, 1993).

Task-function usability. Two dependent measures were used in evaluating the usability of the various in-vehicle tasks as performed under the independent variable conditions tested. The two measures were (1) task completion time and (2) task completion errors. It was hypothesized that drivers experience increased task-function usability with shorter task completion times and/or fewer task errors. Results are presented for each measure when significant main effects or interactions were found. The significance of the results and the implications they may have on ASR system design are also discussed in this section when appropriate.

Task completion time refers to the time duration--in seconds--that it took drivers to complete an in-vehicle task. Time to complete in-vehicle tasks was analyzed in the two mixed-factorial MANOVAs against the independent variables (1) age, input condition, and task type; and (2) age, recognition accuracy, and error type. Two main effects were significant: input condition ($\underline{F}(7,70) = 8.35$, $\underline{p} < 0.001$), and recognition accuracy ($\underline{F}(2,20) = 6.52$, $\underline{p} < 0.010$). The recognition accuracy by error type interaction ($\underline{F}(2,20) = 6.68$, $\underline{p} < 0.010$), and the task type by input condition interaction ($\underline{F}(120,1292) = 11.28$, $\underline{p} < 0.001$) were also significant. Complete ANOVA tables for task completion time are shown in Appendix I, Tables I-17 and I-18.

The significant input condition main effect is shown in Figure 22. A Student Newman-Keuls post-hoc test for the input condition effect showed both 90% accuracy conditions with either rejection or substitution errors and the 75% accuracy with rejection errors condition to be statistically different from the other input conditions. The longest task completion times were found with the 100% accuracy, 75% accuracy with substitution errors, 60% accuracy with either rejection errors or substitution errors, and manual-input conditions. The shortest task completion times are demonstrated under both 90% accuracy conditions with either rejection or substitution errors, and the 75% accuracy with rejection errors condition.

The 100% accuracy condition resulted in the longest average task completion times for any of the input conditions tested (see Figure 22). Drivers always experienced this condition on the first day of testing along with the manual-input condition. A possible explanation for this result is that ASR-controlled tasks are unique and drivers were unaccustomed to performing common in-vehicle tasks using speech input. Command vocabulary and syntax may not have been learned well enough the first day to perform speech-input tasks as efficiently as was demonstrated on subsequent days when accuracy levels were lower. The same explanation may hold for manual-input tasks. None of the participants owned a vehicle identical to the test vehicle. Although subjects were trained on manual tasks, manual-input tasks were performed only on the first day of testing. Therefore, task completion times for the 100% accuracy condition and the manual-input condition may have become shorter with further practice.

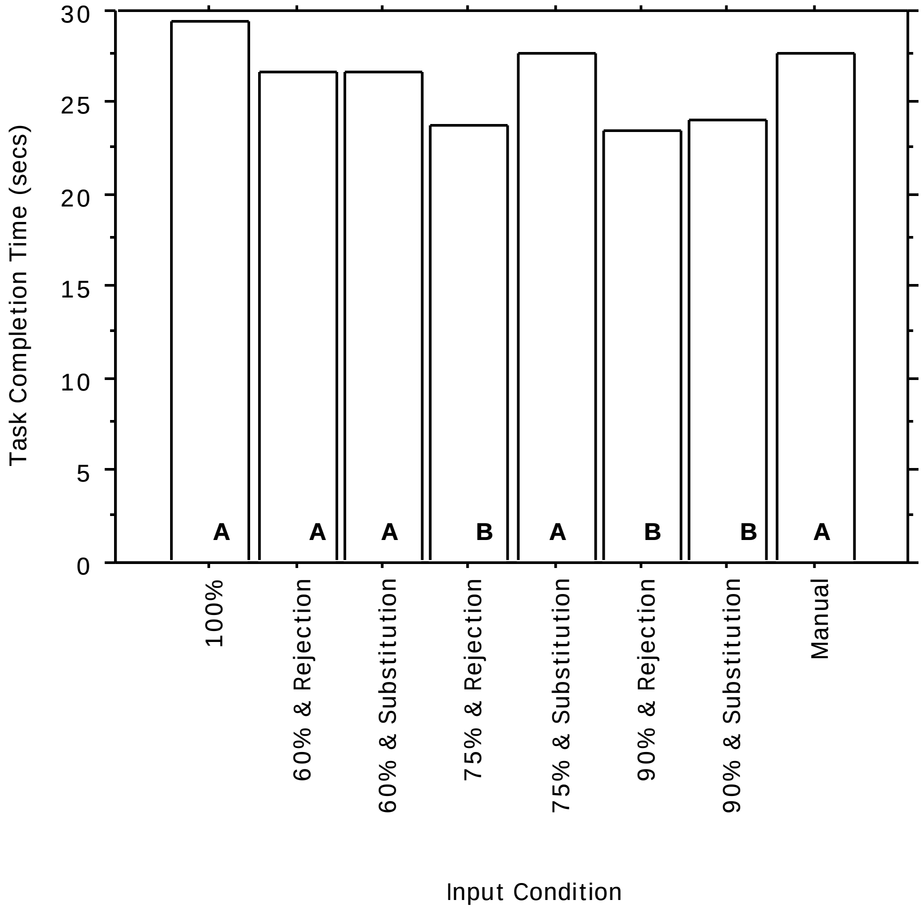

Figure 22. Mean task completion time as a function of input condition. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

The significant recognition accuracy effect is shown in Figure 23. A Student Newman-Keuls post-hoc test showed that the 90% accuracy condition resulted in significantly shorter task completion times compared with the 60% and 75% accuracy conditions. No statistical difference in task completion time was found between the 60% and 75% accuracy conditions.

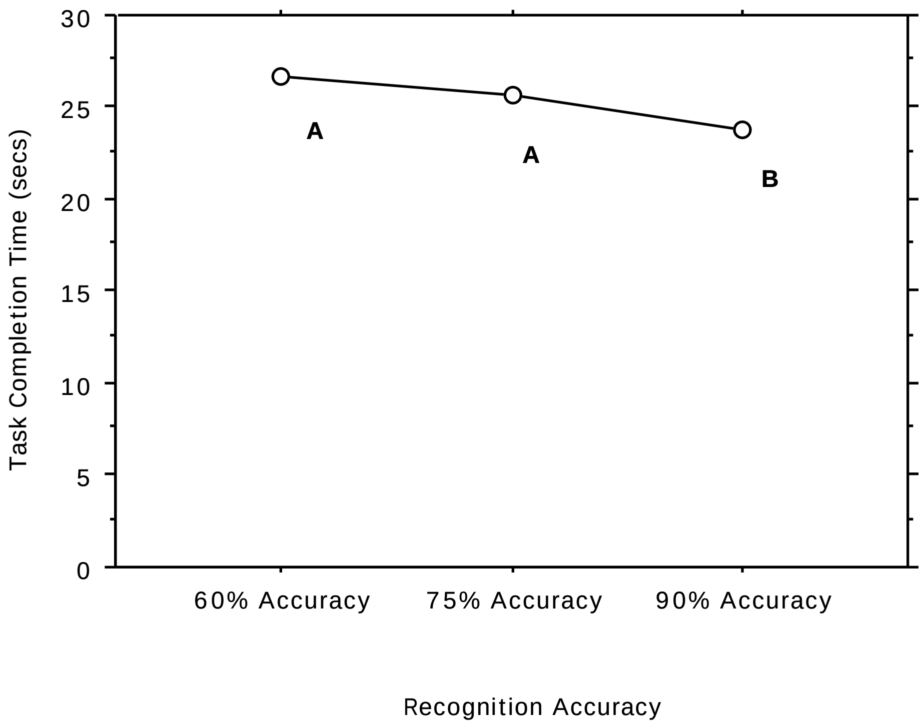

Figure 23. Mean task completion time as a function of recognition accuracy. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

Figure 24 shows the recognition accuracy by error type interaction. The 90% accuracy condition under both error types, and the 75% accuracy condition under the rejection error type resulted in the shortest task completion times. The 75% accuracy condition under the substitution error type showed markedly increased task completion times; greater than even the 60% accuracy condition under both error types.

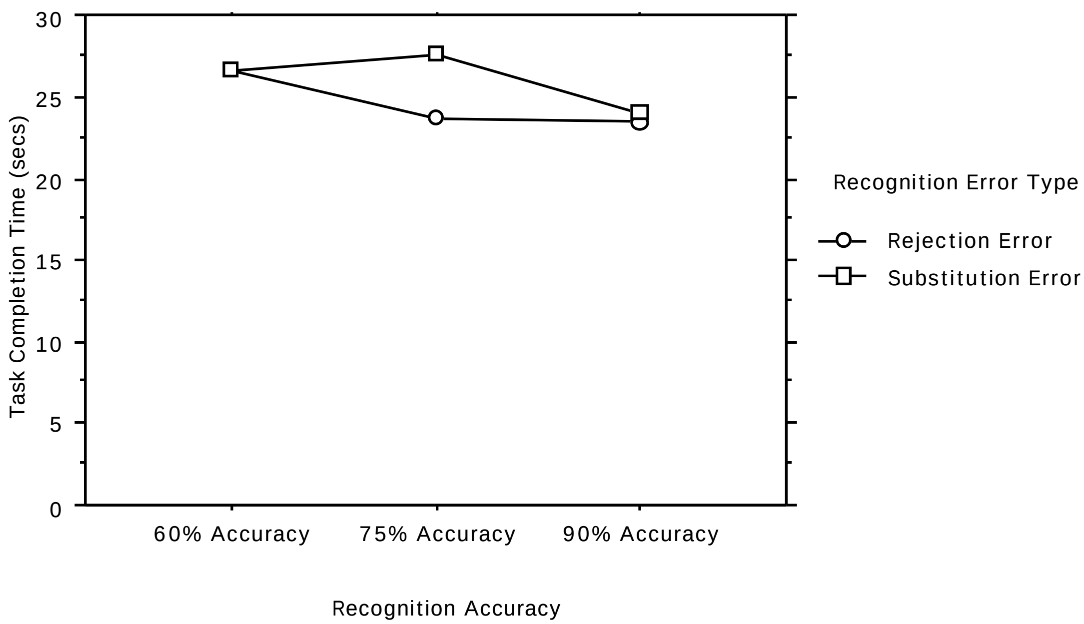

Figure 24. Mean task completion time as a function of recognition accuracy and recognition error type.

Results from the simple effects tests for the task type by input condition interaction are shown in Table 8. In general, the results contradict the overall decision tree recommendations for only three of the task types: (1) activate/deactivate windshield wipers (Task Type number 5); (2) tune radio to specific radio frequency (Task Type number 12); and (3) mute the volume on the radio (Task Type number 14). The results showed that the “windshield wiper” task , the “radio frequency” task, and the “mute volume” task should be performed using speech input instead of the recommended manual input to minimize task completion time. The “radio frequency” task and the “mute volume” task were considered continuous-input type tasks, and because of that consideration the decision trees lead to the recommendation of manual control. Using speech input, the two tasks were actually discrete-input type tasks. Treating a task function as discrete versus continuous can change the recommendations from the decision trees; in particular, the Behavioral Allocation tree. The “windshield wiper” task was recommended for manual input by the decision trees because of its close proximity to the steering wheel and its ease of activation without looking directly at the control. The current decision trees fail to adequately account for such task-function characteristics, leading to recommendations for manual input instead of speech.

The remaining task types had statistically similar results between the manual-input condition and one or more of the speech-input conditions. For these remaining tasks, design tradeoffs for whether the tasks should be performed using speech or manual input could be determined by other criteria such as cost or ease of implementation in place of the decision tree recommendations.

Table 8. Results from the simple effects test for the significant task type by input condition interaction for the mean task completion time dependent measure. (Task Type number corresponds with the numbers for in-vehicle tasks listed described in Table 3 of the Experimental Methodology Section.)

Task Type	Overall Decision Tree Recommendation	Manual	100%	90% w/ Rej. Error	90% w/ Sub. Error	75% w/ Rej. Error	75% w/ Sub. Error	60% w/ Rej. Error	60% w/ Sub. Error
1	Manual	A	A	A	A	A	A	A	A
means		24.11	25.64	22.40	22.44	20.14	23.99	25.70	24.53
2	Speech	A	A	B	B	B	B	B	B
means		28.65	28.96	18.78	15.85	19.26	18.46	22.25	17.98
3	Manual	A	A	A	A	A	A	A	A
means		28.4	32.39	30.66	26.91	31.40	34.20	31.58	34.64
4	Manual	A	A	A	A	A	A	A	A
means		26.92	31.65	27.76	31.10	25.46	34.11	26.53	27.24
5	Manual	BC	ABC	CD	BC	CD	A	D	AB
means		27.28	30.07	24.27	27.75	24.27	35.59	20.26	32.36
6	Manual	BCD	AB	D	CD	CD	ABC	A	CD
means		25.95	33.69	18.82	22.82	22.76	29.44	35.77	24.64
7	Speech	C	AB	BC	AB	C	AB	ABC	A
means		23.79	35.13	27.40	35.72	23.97	36.36	29.91	37.86
8	Speech	A	A	A	A	A	A	A	A
means		24.19	25.52	18.22	16.42	23.39	20.79	20.20	16.55
9	Speech	AB	AB	AB	B	AB	AB	AB	A
means		33.69	28.29	27.79	23.89	30.92	33.42	32.26	36.27
10	Speech	A	AB	ABC	ABC	BC	C	BC	BC
means		32.87	32.29	28.00	29.19	24.21	23.47	24.48	24.64
11	Speech	A	A	A	A	A	A	A	A
means		18.75	22.09	18.35	21.68	15.31	21.50	20.08	21.74
12	Manual	A	B	B	B	B	B	B	B
means		29.15	23.77	19.31	18.72	18.97	19.24	19.49	18.06
13	Speech	A	A	A	A	A	A	A	A
means		21.91	22.19	19.05	17.85	17.49	18.65	23.57	19.44
14	Manual	AB	A	C	C	BC	AB	AB	AB
means		24.35	28.01	15.13	15.19	17.36	23.31	22.80	23.63
15	Speech	A	AB	B	AB	AB	AB	AB	AB
means		35.03	32.09	24.02	24.79	26.40	28.05	25.38	32.16
16	Speech	CD	CD	D	CD	CD	ABC	AB	A
means		37.77	42.61	36.20	39.94	39.23	46.18	48.84	52.43

Task completion errors were recorded whenever drivers did not correctly complete a speech-input task. To analyze task errors, either a Chi-square test, a McNemar change test, or a Cochran *Q* test (Siegel and Castellan, 1988) was performed to examine the effects of age, input condition, recognition accuracy, error type, and type of in-vehicle task on correct task completion by drivers. A Chi-square test was performed on the independent variable, age. The McNemar change test was performed on the error type independent variable. The Cochran *Q* test was performed on the input condition, the recognition accuracy, and the task-type independent variables. The type of test used was determined by the number of levels for the independent variable examined, and whether the data was from independent or related samples. These nonparametric tests determined the significance of differences between two or more related, or independent, groups (i.e., the levels of the independent variables) with respect to the number of task errors that occurred. It was appropriate to use nonparametric statistics for this data because it was only categorical-type data.

The results showed an effect of age ($X^2(1,\underline{N}=4608) = 21.479$, $\underline{p} < 0.001$), input condition ($Q\ (7,\underline{N}=4608) = 294.674$, $\underline{p} < 0.001$), recognition accuracy ($Q\ (3,\underline{N}=4032) = 23.380$, $\underline{p} < 0.001$), error type ($X^2(1,\underline{N}=3456) = 102.030$, $\underline{p} < 0.001$), and task-type ($Q\ (15,\underline{N}=4608) = 1397.588$, $\underline{p} < 0.001$) on correct completion of in-vehicle tasks. Older drivers accounted for the significant proportion of task errors recorded (79 errors out of a total 111 recorded errors). The 75% and 60% accuracy with substitution error conditions produced the significant proportion of task errors recorded (42 errors and 55 errors out of a total 111 recorded errors, respectively). The 75% accuracy and 60% accuracy conditions collapsed across age and error type resulted in the significant proportion of task errors recorded (42 errors and 54 errors out of 107 recorded errors, respectively). The substitution error condition accounted for the significant proportion of task errors recorded when collapsed across age and recognition accuracy (105 errors out of a total 111 recorded errors). The average fuel economy display task, the "seek-up" the radio stations task, the cellular phone preset number task, and the cellular phone personal number task account for the significant proportion of task errors recorded collapsed across all the variables except task type (17, 21, 32, and 21 errors out of a total 111 recorded errors, respectively).

Task errors that were recorded in the analysis for the task type independent variable occurred only when a speech input condition was employed, and the accuracy levels were either 75% or 60% with substitution errors. Therefore, these results showed that all the tasks should be performed using either manual input, or speech input with high levels of recognition accuracy (e.g., 90% or greater) and/or with ASR systems that produce only rejection errors. Criteria such as cost or ease of implementation could be used in place of the decision tree recommendations to determine whether the tasks should be performed using speech or manual input.

Driver preference/acceptance of task functions. Seven-point Likert-type subjective questionnaires were used to evaluate driver preference/acceptance of in-vehicle tasks (see Appendix G). The questionnaires measured drivers' perceived "ease-of-use," "comfort level," and "distraction level" experienced while performing the various in-vehicle tasks. Results are presented for each measure when significant main effects or interactions were found. The significance of the results and the implications they may have on ASR system design are also discussed in this section when appropriate.

The first question a driver was asked after completing a task was, "Using the seven-point ease-of-use scale, how easy was it to perform this task while driving the vehicle?" The results from the "ease-of-use" question were analyzed in the two mixed-factorial MANOVAs against the independent variables (1) age, input condition, and task type; and (2) age, recognition accuracy, and error type. Two main effects were found significant: input condition ($\underline{F}(7,70) = 4.19$, $\underline{p} < 0.001$) and recognition accuracy

($\underline{F}(2,20) = 9.89$, $\underline{p} < 0.001$). The task type by input condition interaction was also significant ($\underline{F}(120,1292) = 6.23$, $\underline{p} < 0.001$). Complete ANOVA tables for ease-of-use ratings are shown in Appendix I, Table I-19 and I-20.

The significant input condition and recognition accuracy main effects are shown in Figures 25 and 26, respectively. A Student Newman-Keuls post-hoc test for the input condition effect found no statistical differences between the 60% accuracy conditions with either rejection errors or substitution errors, and the manual-input condition. The 60% accuracy with substitution errors condition was statistically different from the 100% accuracy input condition, and both 90% accuracy conditions with either rejection errors or substitution errors, and resulted in the lowest ease-of-use ratings (see Figure 25).

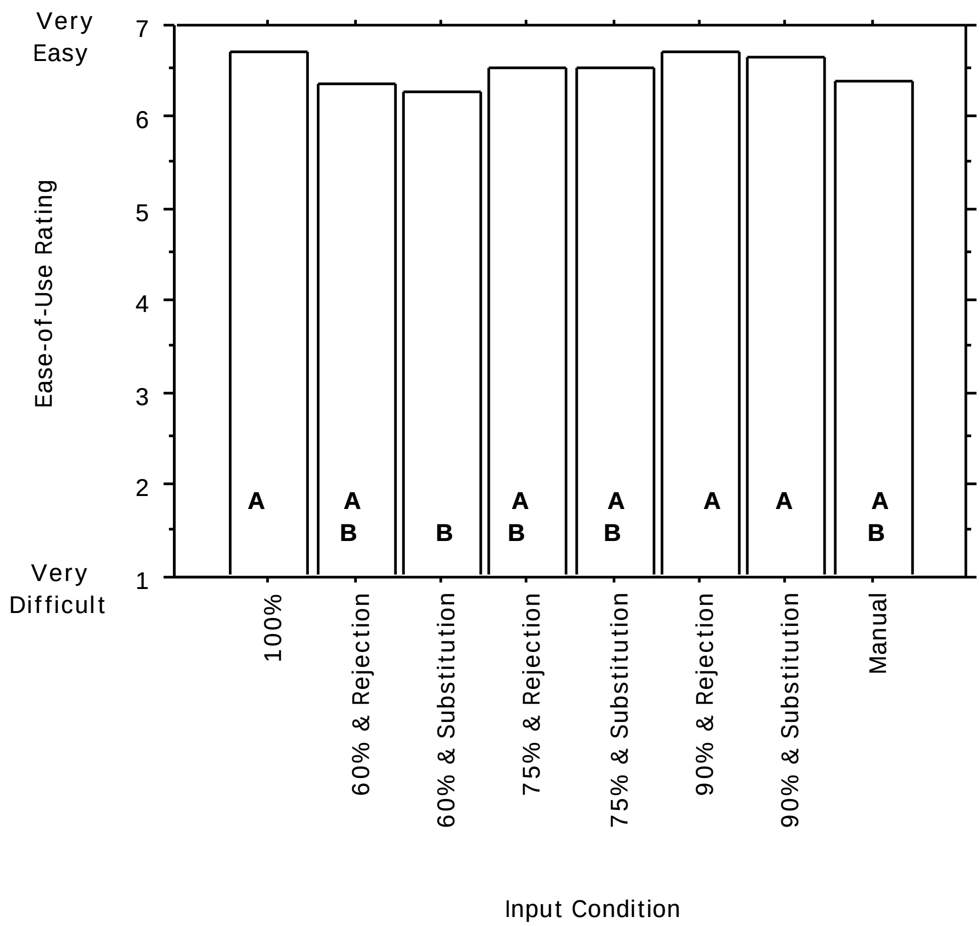

Figure 25. Mean ease-of-use rating as a function of input condition. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

A Student Newman-Keuls post-hoc test for the recognition accuracy effect showed that both the 90% and 75% recognition accuracy conditions result in statistically higher ease-of-use ratings than the 60% recognition accuracy condition (see Figure 26). No statistical difference was found between the 90% and 75% recognition accuracy conditions.

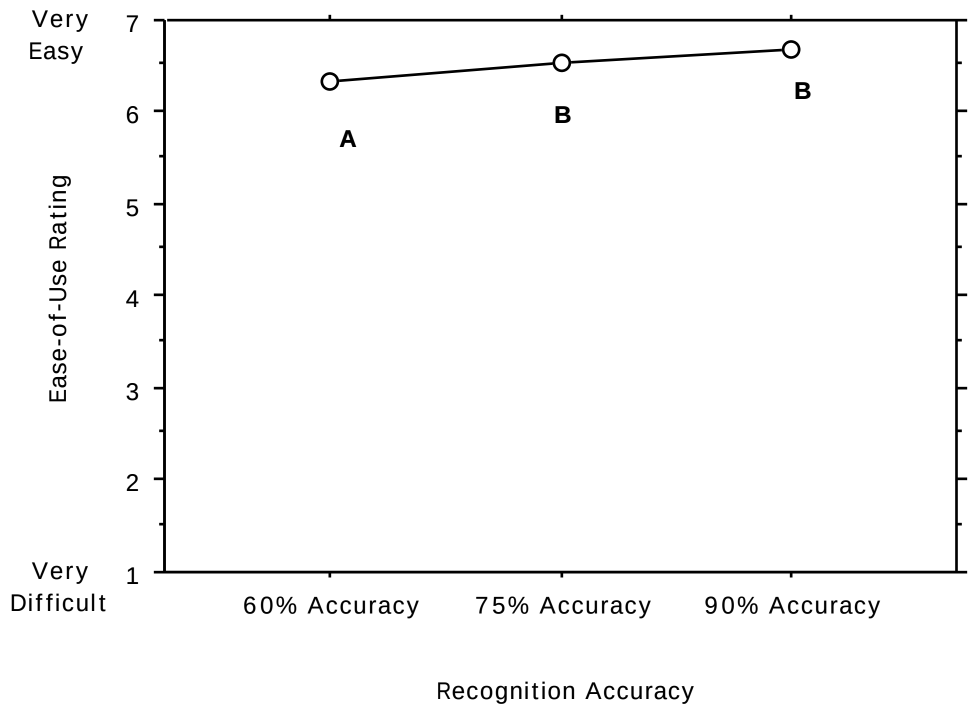

Figure 26. Mean ease-of-use rating as a function of recognition accuracy. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

Results from the simple effects tests for the task type by input condition interaction are shown in Table 9. In general, the results contradict the overall decision tree recommendations for only one of the task types: tune radio to specific radio frequency (Task Type number 12). The results showed that the “radio frequency” task should be performed using speech input instead of the recommended manual input to increase subjective ratings for ease-of-use. The “radio frequency” task was considered a continuous-input type task, and because of that consideration the decision trees lead to the recommendation of manual control. Using speech input, the task was actually a discrete-input type task. Treating a task function as discrete versus continuous can change the recommendations from the decision trees; in particular, the Behavioral Allocation tree. The remaining task types had statistically similar results between the manual-input condition and one or more of the speech-input conditions. For these remaining tasks, design tradeoffs for whether the tasks should be performed using speech or manual input could be determined by other criteria such as cost or ease of implementation in place of the decision tree recommendations.

Table 9. Results from the simple effects test for the significant task type by input condition interaction for the Ease-of-Use subjective rating dependent measure. (Task Type number corresponds with the numbers for in-vehicle tasks listed described in Table 3 of the Experimental Methodology Section.)

Task Type	**Overall Decision Tree Recommendation**	**Manual**	**100%**	**90% w/ Rej. Error**	**90% w/ Sub. Error**	**75% w/ Rej. Error**	**75% w/ Sub. Error**	**60% w/ Rej. Error**	**60% w/ Sub. Error**
1	Manual	A	A	A	A	AB	A	B	B
means		6.72	6.89	6.78	6.72	6.58	6.67	6.33	6.36
2	Speech	B	A	A	A	AB	A	AB	AB
means		6.33	6.72	6.81	6.83	6.56	6.69	6.64	6.61
3	Manual	AB	A	AB	AB	AB	B	B	B
means		6.50	6.86	6.69	6.64	6.64	6.42	6.39	6.33
4	Manual	A	A	A	A	AB	AB	AB	B
means		6.78	6.75	6.67	6.69	6.50	6.53	6.42	6.19
5	Manual	A	A	A	A	A	A	A	B
means		6.67	6.83	6.81	6.69	6.75	6.42	6.89	5.89
6	Manual	AB	A	A	A	A	AB	B	AB
means		6.56	6.86	6.83	6.83	6.69	6.5	6.31	6.47
7	Speech	A	A	A	AB	A	AB	AB	B
means		6.67	6.42	6.47	6.19	6.36	6.14	6.17	5.58
8	Speech	A	A	A	A	AB	A	AB	B
means		6.63	6.83	6.83	6.83	6.50	6.72	6.50	6.30
9	Speech	C	ABC	A	A	AB	A	ABC	BC
means		6.05	6.48	6.72	6.69	6.52	6.58	6.44	6.10
10	Speech	C	AB	A	AB	AB	AB	AB	BC
means		6.08	6.45	6.72	6.63	6.61	6.52	6.55	6.27
11	Speech	B	A	A	AB	AB	A	AB	B
means		6.44	6.80	6.80	6.63	6.69	6.80	6.50	6.41
12	Manual	B	A	A	A	A	A	A	A
means		5.97	6.83	6.69	6.75	6.47	6.83	6.72	6.58
13	Speech	A	A	A	A	A	A	B	A
means		6.66	6.83	6.69	6.83	6.72	6.83	6.30	6.56
14	Manual	AB	A	AB	AB	AB	B	AB	AB
means		6.75	6.88	6.83	6.83	6.83	6.50	6.69	6.76
15	Speech	BC	AB	A	AB	AB	ABC	ABC	C
means		5.68	6.36	6.61	6.30	6.25	6.08	6.11	5.52
16	Speech	A	A	A	A	A	A	A	A
means		5.86	6.20	6.24	5.95	6.08	6.37	5.36	5.61

The second question a driver was asked after completing a task was, "Using the seven-point comfort scale, how comfortable did you feel performing this task while driving the vehicle?" The results from the "comfort" question were analyzed in the two mixed-factorial MANOVAs against the independent variables: (1) age, input condition, and task type; and (2) age, recognition accuracy, and error type. Two main effects were found significant: input condition ($F(7,70) = 4.86$, $p < 0.001$) and recognition accuracy ($F(2,20) = 10.52$, $p < 0.001$). The task type by input condition interaction was also significant ($F(120,1292) = 6.88$, $p < 0.001$). Complete ANOVA tables for comfort ratings are shown in Appendix I, Table I-21 and I-22.

The significant input condition and recognition accuracy main effects are shown in Figures 27 and 28, respectively. A Student Newman-Keuls post-hoc test for the input condition effect found statistical differences between the 90% accuracy with rejection error condition, and both the manual-input and 60% accuracy with substitution errors conditions. No statistical difference was found between the 60% accuracy conditions with either rejection errors or substitution errors, and the manual-input conditions (see Figure 27). The 60% accuracy with substitution errors condition resulted in the lowest comfort ratings of all the input conditions tested.

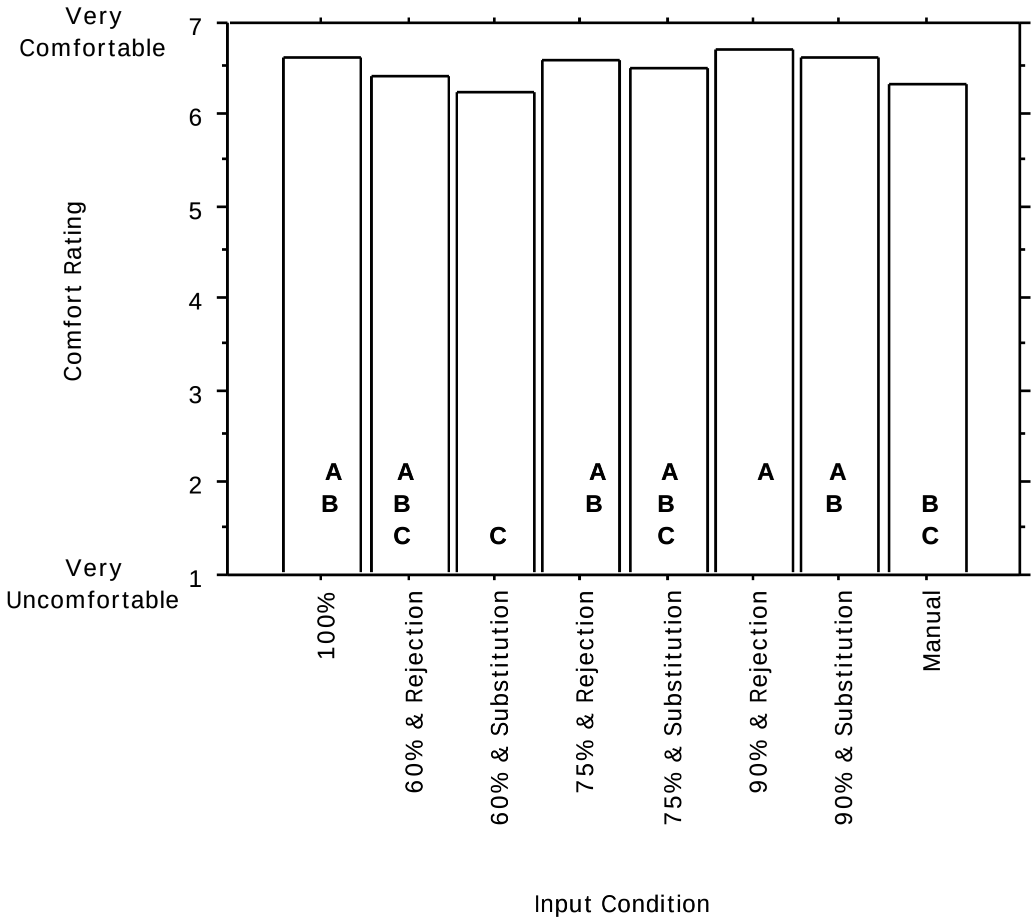

Figure 27. Mean comfort rating as a function of input condition. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

A Student Newman-Keuls post-hoc test for the recognition accuracy effect shows that both the 90% and 75% recognition accuracy conditions resulted in statistically higher comfort ratings than the 60% recognition accuracy condition (see Figure 28). No statistical difference was found between the 90% and 75% recognition accuracy conditions.

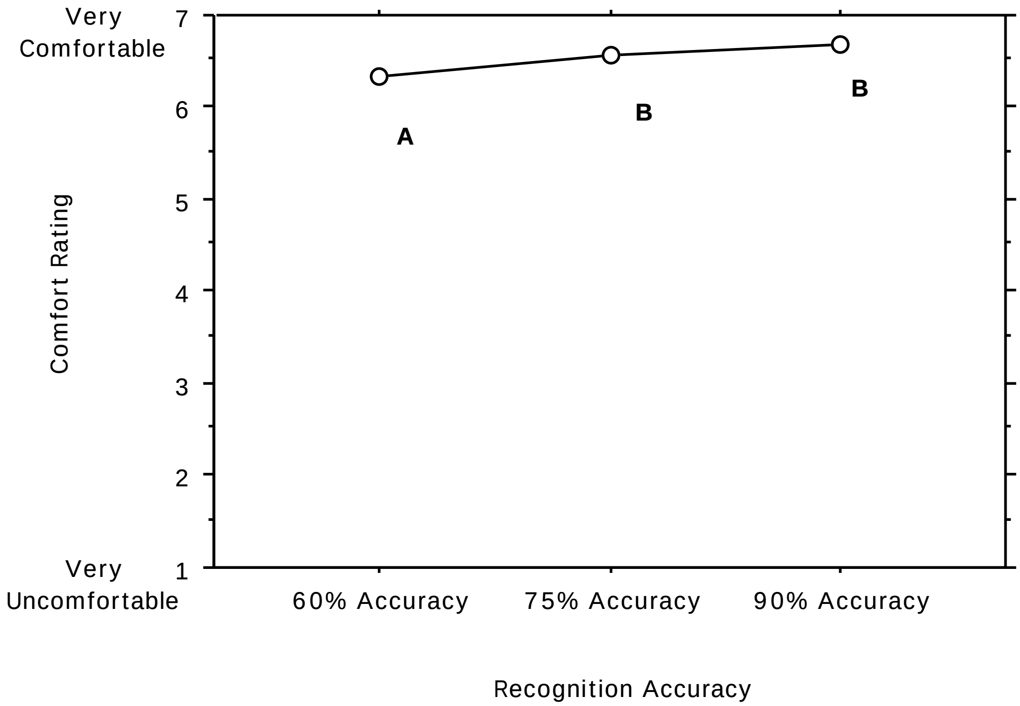

Figure 28. Mean comfort rating as a function of recognition accuracy. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

Results from the simple effects tests for the task type by input condition interaction are shown in Table 10. In general, the results contradict the overall decision tree recommendations for only one of the task types: tune radio to specific radio frequency (Task Type number 12). The results showed that the “radio frequency” task should be performed using speech input instead of the recommended manual input to increase subjective ratings for comfort. The “radio frequency” task was considered a continuous-input type task, and because of that consideration the decision trees lead to the recommendation of manual control. Using speech input, the task was actually a discrete-input type task. As stated earlier, treating a task function as discrete versus continuous can change the recommendations from the decision trees; in particular, the Behavioral Allocation tree. The remaining task types had statistically similar results between the manual-input condition and one or more of the speech-input conditions. For these remaining tasks, design tradeoffs for whether the tasks should be performed using speech or manual input could be determined by other criteria such as cost or ease of implementation in place of the decision tree recommendations.

Table 10. Results from the simple effects test for the significant task type by input condition interaction for the Comfort subjective rating dependent measure. (Task Type number corresponds with the numbers for in-vehicle tasks listed described in Table 3 of the Experimental Methodology Section.)

Task Type	Overall Decision Tree Recommendation	Manual	100%	90% w/ Rej. Error	90% w/ Sub. Error	75% w/ Rej. Error	75% w/ Sub. Error	60% w/ Rej. Error	60% w/ Sub. Error
1	Manual	ABC	A	AB	AB	ABC	ABC	BC	C
means		6.67	6.83	6.78	6.75	6.61	6.69	6.47	6.39
2	Speech	B	AB	AB	A	AB	AB	AB	AB
means		6.44	6.67	6.81	6.83	6.64	6.67	6.64	6.58
3	Manual	AB	A	A	A	A	AB	AB	B
means		6.53	6.78	6.75	6.64	6.69	6.40	6.56	6.22
4	Manual	A	A	A	A	A	A	A	B
means		6.83	6.64	6.64	6.56	6.53	6.47	6.50	6.19
5	Manual	A	A	A	A	A	A	A	B
means		6.74	6.81	6.81	6.69	6.81	6.47	6.58	5.92
6	Manual	A	A	A	A	A	A	A	A
means		6.53	6.83	6.83	6.83	6.75	6.42	6.44	6.44
7	Speech	A	AB	AB	AB	AB	AB	AB	B
means		6.61	6.33	6.50	6.22	6.47	6.19	6.03	5.75
8	Speech	DC	AB	A	A	ABC	ABC	BCD	D
means		6.50	6.77	6.80	6.83	6.61	6.72	6.52	6.33
9	Speech	C	AB	A	A	A	A	AB	BC
means		5.83	6.30	6.66	6.63	6.55	6.54	6.44	6.10
10	Speech	C	AB	A	AB	A	AB	AB	B
means		5.94	6.36	6.72	6.55	6.69	6.51	6.52	6.25
11	Speech	BC	A	AB	ABC	ABC	AB	ABC	C
means		6.47	6.83	6.80	6.66	6.66	6.77	6.55	6.41
12	Manual	B	A	A	A	A	A	A	A
means		5.80	6.75	6.66	6.75	6.58	6.80	6.72	6.55
13	Speech	A	A	AB	A	AB	A	B	AB
means		6.58	6.83	6.72	6.80	6.69	6.83	6.50	6.61
14	Manual	A	A	A	A	A	B	AB	B
means		6.77	6.86	6.83	6.83	6.83	6.47	6.69	6.48
15	Speech	D	AB	A	AB	AB	BCD	ABC	CD
means		5.51	6.25	6.61	6.33	6.38	5.88	6.11	5.61
16	Speech	A	A	A	A	A	A	A	A
means		5.75	6.00	6.33	5.95	6.05	6.14	5.42	5.50

The final question a driver was asked after completing a task was, "Using the seven-point distraction scale, how distracting was performing this task while driving the vehicle?" The results from the "distraction" question were analyzed in the two mixed-factorial MANOVAs against the independent variables: (1) age, input condition, and task type; and (2) age, recognition accuracy, and error type. Three main effects were found significant: input condition ($F(7,70) = 8.37$, $p < 0.001$), recognition accuracy ($F(2,20) = 20.93$, $p < 0.001$), and error type ($F(1,10) = 14.34$, $p < 0.005$). The task type by input condition interaction was also significant ($F(120,1292) = 12.86$, $p < 0.001$). Complete ANOVA tables for distraction rating are shown in Appendix I, Tables I-23 and I-24.

The significant input condition main effect is shown in Figure 29. A Student Newman-Keuls post-hoc test for the input condition effect found the manual-input condition statistically similar in distraction rating to both 60% accuracy conditions with either rejection errors or substitution errors, and the 75% accuracy with substitution error condition. The remaining input conditions resulted in statistically lower ratings of distraction than the manual-input condition and the 60% accuracy with substitution errors condition (see Figure 29). It is important to note that larger distraction rating numbers mean the task was less distracting to perform.

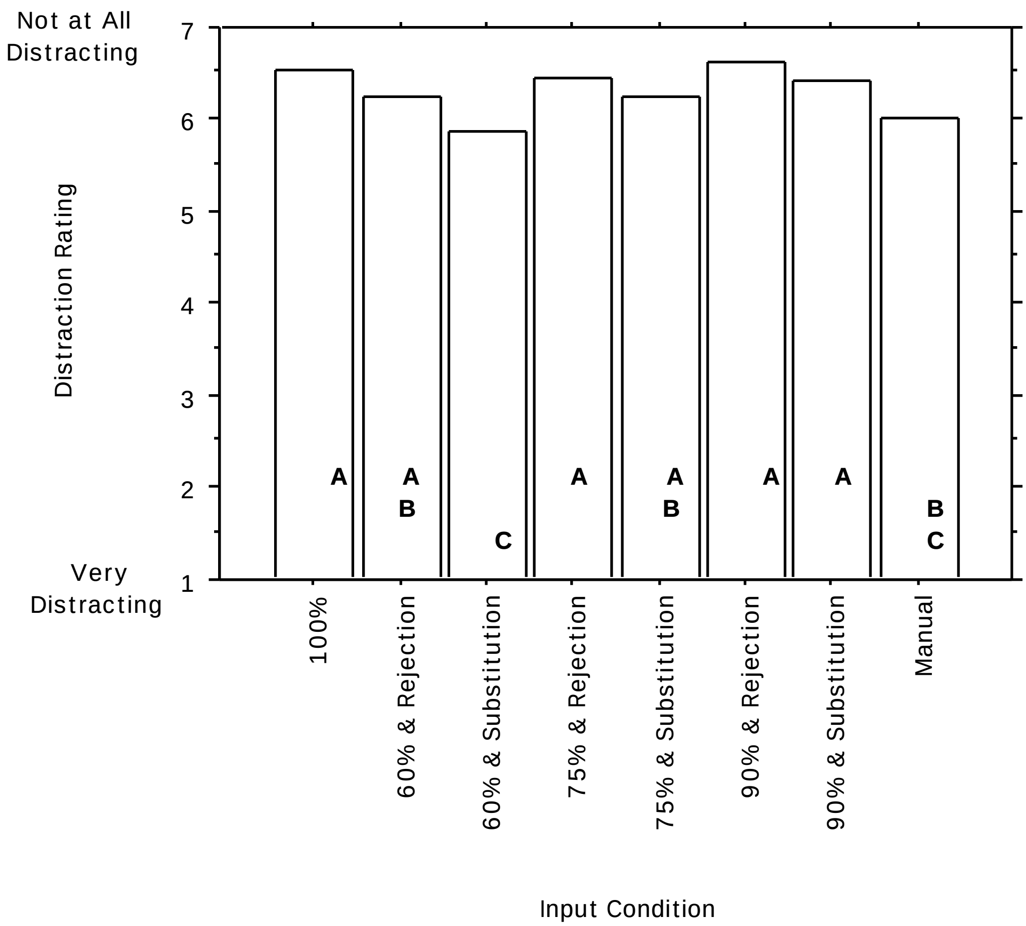

Figure 29. Mean distraction rating as a function of input condition. Higher rating numbers indicate less distraction. (Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.)

Figure 30 shows the significant recognition accuracy main effect. The 90% accuracy condition resulted in the lowest subjective ratings of distraction, followed by the 75% accuracy condition, and next the 60% accuracy condition (Note: larger distraction rating numbers mean the task was less distracting to perform). A Student Newman-Keuls post-hoc test for the recognition accuracy effect found statistical differences between all three accuracy conditions.

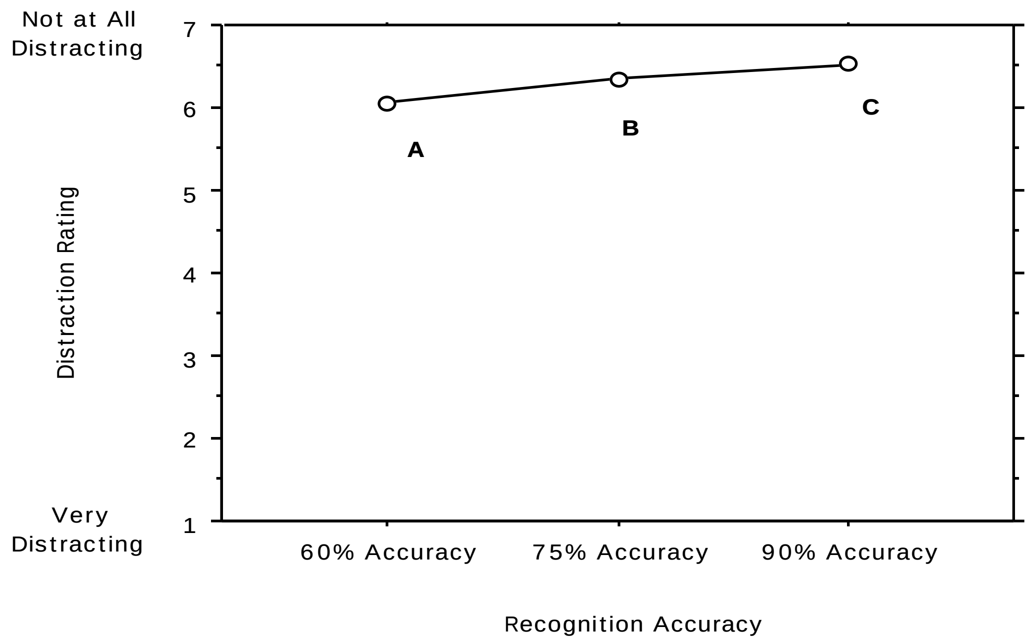

Figure 30. Mean distraction rating as a function of recognition accuracy. Higher rating numbers indicate less distraction. Bold letters indicate significant differences in treatments. Treatments not sharing a common letter differ significantly from each other.

The significant error type main effect showed that the rejection error condition resulted in statistically lower subjective ratings of distraction than the substitution error condition. (Mean distraction rating for rejection error condition = 6.44; mean distraction rating for substitution error condition = 6.19. Note: larger distraction rating numbers mean the task was less distracting to perform.) Drivers may have rated tasks that contained a rejection error as being "less" distracting (higher distraction rating scale numbers) because the ASR system alerted them through visual plus auditory feedback that such an error had occurred. Substitution errors during task performance required the driver to detect and correct the error without assistance from the ASR system.

Results from the simple effects tests for the task type by input condition interaction are shown in Table 11. In general, the results contradict the overall decision tree recommendations for only two of the task types: (1) adjust temperature control (Task Type Number 6); and (2) tune radio to specific radio frequency (Task Type number 12). The results showed that the “adjust temperature” task and the “radio frequency” task should be performed using speech input instead of the recommended manual input to improve subjective ratings for distraction. The “adjust temperature” task and the “radio frequency” task were considered continuous-input type tasks, and because of that consideration the decision trees lead to the recommendation of manual control. Using speech input, the two tasks were actually discrete-input type tasks. Treating a task function as discrete versus continuous can change the recommendations from the decision trees; in particular, the Behavioral Allocation tree. The remaining task types had statistically similar results between the manual-input condition and one or more of the speech-input conditions. For these remaining tasks, design tradeoffs for whether the tasks should be performed using speech or manual input could be determined by other criteria such as cost or ease of implementation in place of the decision tree recommendations.

Table 11. Results from the simple effects test for the significant task type by input condition interaction for the Discomfort subjective rating dependent measure. (Task Type number corresponds with the numbers for in-vehicle tasks listed described in Table 3 of the Experimental Methodology Section.)

Task Type	Overall Decision Tree Recommendation	Manual	100%	90% w/ Rej. Error	90% w/ Sub. Error	75% w/ Rej. Error	75% w/ Sub. Error	60% w/ Rej. Error	60% w/ Sub. Error
1	Manual	AB	A	A	AB	AB	AB	BC	C
means		6.58	6.89	6.81	6.58	6.58	6.56	6.33	6.14
2	Speech	B	AB	A	A	AB	AB	AB	AB
means		6.12	6.64	6.83	6.81	6.47	6.61	6.61	6.50
3	Manual	ABC	A	A	AB	AB	BC	AB	C
means		6.22	6.75	6.75	6.47	6.56	6.07	6.39	5.86
4	Manual	A	AB	AB	B	B	B	B	C
means		6.89	6.58	6.64	6.44	6.53	6.35	6.50	6.00
5	Manual	A	A	A	AB	A	B	AB	C
means		6.68	6.81	6.81	6.58	6.75	6.28	6.50	5.53
6	Manual	B	A	A	A	A	B	B	B
means		6.08	6.67	6.75	6.64	6.64	6.08	6.19	6.17
7	Speech	A	AB	AB	AB	AB	B	AB	C
means		6.42	5.92	6.19	5.78	6.19	5.53	5.81	4.72
8	Speech	BC	A	A	A	AB	AB	AB	C
means		6.19	6.75	6.72	6.72	6.44	6.50	6.44	6.02
9	Speech	B	A	A	A	A	A	A	B
means		5.33	6.21	6.50	6.30	6.44	6.23	6.22	5.33
10	Speech	C	AB	A	AB	A	AB	AB	BC
means		5.61	6.15	6.52	6.27	6.52	6.12	6.19	5.87
11	Speech	C	A	A	ABC	AB	A	ABC	BC
means		6.02	6.80	6.77	6.38	6.55	6.75	6.41	6.13
12	Manual	B	A	A	A	A	A	A	A
means		5.19	6.69	6.58	6.66	6.38	6.80	6.69	6.27
13	Speech	B	A	AB	AB	AB	AB	B	AB
means		6.25	6.80	6.66	6.69	6.58	6.62	6.22	6.40
14	Manual	AB	A	B	A	A	B	AB	B
means		6.66	6.86	6.83	6.77	6.83	6.36	6.63	6.37
15	Speech	C	AB	A	AB	AB	B	B	C
means		4.84	6.08	6.47	5.94	6.02	5.55	5.75	4.85
16	Speech	AB	A	A	AB	A	AB	AB	B
means		5.08	5.83	6.06	5.41	5.66	5.44	5.03	4.50

The results from the measures of driver preference/acceptance showed an inter-correlation between the subjective constructs. This was similar to results found in Experiment 1. Again, this suggests that all three dependent measures (ease-of-use, comfort, and distraction) were assessing the same subjective construct. There may have been only a single construct that was evaluated by subjects when performing an in-vehicle task. Future research could employ only one of the of the subjective measures since it appears that ease-of-use, comfort, and distraction are all correlated. The distraction question is offered as the choice for evaluating measures of driver preference/acceptance.

It interesting to note the similarity in results for measures of driver preference/acceptance between Experiments 1 and 2. For each experiment, higher subjective ratings were associated with higher levels of system performance. However, the average subjective ratings never fell below the mid point of the scale (i.e., average subjective ratings never fell on the negative end of the subjective scales) despite experimental conditions that demonstrated rather low levels of ASR system performance.

Summary of Results and Discussion

This section summarizes the significant results from the two on-road experiments, and attempts to attach further importance to the numbers presented above through recommendations for ASR system design. The section is divided into the four ASR system parameters evaluated in the experiments (recognition errors, recognition accuracy, ASR system feedback, and error correction method), and the influence that input-condition and driver age had on the dependent measures recorded. A further understanding regarding the effects that ASR system design parameters had on the criteria of interest, compared to current manually-controlled systems, is presented.

First, summary tables of the significant results from Experiments 1 and 2 are presented in Tables 12 and 13, respectively. These tables show all significant main effects and interactions from the multivariate analyses of variance performed on the dependent measures recorded in the two experiments.

Table 12. Summary table of significant results from Experiment 1 (p -values < 0.10).

Dependent Measure	**Significant Main Effects**	**p - values**	**Significant Interactions**	**p - values**
Steering Variance	Error Type	0.0009		
Steering Reversals > 6°				
Steering Reversals > 12°				
Accelerator Position Variance	Error Type	0.0013		
Mean Vehicle Velocity				
Vehicle Velocity Variance	Error Type	0.0001		
Brake Activations	Error Type	0.0015	Feedback * Error Type	0.0922
Reaction Time to Forward Scene Event				
Task Completion Time	Error Type	0.0001	Error Correction * Error Type	0.0029
Ease-of-Use Rating	Error Type	0.0001	Error Correction * Error Type	0.0390
Comfort Rating	Error Type	0.0002		
Distraction Rating	Error Type	0.0001	Feedback * Error Correction	0.0724
			Feedback * Error Type	0.0970

Table 13. Summary table of significant results from Experiment 2 (p -values < 0.05).

Dependent Measure	Significant Main Effects	p - values	Significant Interactions	p - values
Percent Eyes-Off-Road	Input Condition	0.0001	Task Type * Input Condition	0.0093
	Accuracy	0.0076		
Steering Variance	Age	0.0001		
Steering Reversals > 6°	Age	0.0361	Task Type * Input Condition	0.0001
	Input Condition	0.0014		
	Accuracy	0.0002		
Steering Reversals > 12°	Age	0.0123	Task Type * Input Condition	0.0003
	Input Condition	0.0098		
	Accuracy	0.0006		
Accelerator Position Variance	Age	0.0001		
Mean Vehicle Velocity				
Vehicle Velocity Variance				
Brake Activations	Age	0.0001	Age * Input Condition	0.0379
			Task Type * Input Condition	0.0001
Task Completion Time	Input Condition	0.0001	Accuracy * Error Type	0.0060
	Accuracy	0.0066	Task Type * Input Condition	0.0001

Table 13 (continued). Summary table of significant results from Experiment 2 (p -values < 0.05).

Dependent Measure	**Significant Main Effects**	**p - values**	**Significant Interactions**	**p - values**
Ease-of-Use Rating	Input Condition	0.0007	Task Type * Input Condition	0.0001
	Accuracy	0.0010		
Comfort Rating	Input Condition	0.0002	Task Type * Input Condition	0.0001
	Accuracy	0.0008		
Distraction Rating	Input Condition	0.0001	Task Type * Input Condition	0.0001
	Accuracy	0.0001		
	Error Type	0.0036		

Recognition errors. The ASR system parameter that affected driving performance most was the type of recognition error allowed by the system. The greatest negative impact to driving performance was evident when substitution errors occurred during task performance. Steering-input behavior was degraded most by substitution errors (see Figures 2), as was speed-maintenance behavior (see Figures 3, 4, and 5). However, no statistical differences were found between the no error condition and tasks that contained rejection errors for the same measures of steering-input behavior and speed maintenance behavior (see Figures 2, 3, 4, and 5). Degradation in driving performance caused by substitution errors, although statistically significant, was relatively small. For example, the difference between the no error condition and the substitution error condition for steering variability was approximately 0.10 degrees of steering wheel rotation (see Figure 2); the difference in number of brake activations was less than 1 activation per minute (see Figure 5); and the difference in speed variability was approximately 3 mph (see Figure 4). This is indicative that recognition errors have a relatively small practical effect on driving performance. From an automotive systems design perspective, ASR systems with recognition errors that are predominantly substitutions should not be eliminated from consideration based only on the effect these systems may have on driving performance.

Measures of task-function usability were also significantly affected by the type of recognition error that occurred during task performance. Both rejection errors and substitution errors demonstrated significantly longer task completion times compared with a no error condition (see Figure 7). However, no difference in task completion times was found between rejection errors or substitution errors when drivers performed speech-input tasks. Differences in task completion times between a no error condition and a recognition error condition averaged approximately 2 seconds across all in-vehicle tasks tested.

The number of task completion errors was also significantly influenced by recognition error type. A task completion error was recorded whenever the driver did not correctly complete an instructed task. Substitution errors accounted for the large majority task completion errors recorded in Experiments 1 and 2 (285 errors out of a total 293 recorded errors for both experiments). Rejection errors did not lead to a single task completion error recorded for either experiment.

Finally, driver preference/acceptance of ASR-controlled tasks was negatively affected by the type of recognition error that occurred during the in-vehicle task interaction. Subjective ratings of usability, comfort, and distraction while performing speech-input tasks were lowest when substitution errors occurred, followed next by rejection errors (see Figures 9, 11, and 12). A statistical difference was found between all three error type conditions: no error, rejection error, and substitution error. However, the difference in subjective ratings between error conditions was rather small. For both the five-point and seven-point Likert-type scales, the difference in ratings between error conditions (i.e., between no error, rejection error, and substitution error) was less than one.

The following recommendations are made regarding the type(s) of recognition errors allowable by an automotive ASR system. An automotive ASR system should minimize the number of recognition errors that occur during interactions with the system. If the system is unable to perform with 100% accuracy, then either rejection errors or substitution errors are acceptable when the system performs with accuracy levels greater than 90% (see Figures 17, 19, 22, 25, 27, and 28). However, rejection errors are preferred over substitution errors when a system performs with accuracy levels as low as 75% (see Figures 22 and 24).

Recognition accuracy. The results from Experiment 2 show that recognition accuracy significantly affected driving performance, task-function usability, and driver preference/acceptance. Recognition accuracy levels of 90% and 60% resulted in statistically lower percent eyes-off-road times than the 75% recognition accuracy level (see Figure 16). No statistical difference was found between the 90% and 60% accuracy conditions. Differences between the 75% accuracy level and both the 60% and 90% accuracy levels in mean percent eyes-off-road during speech-input task performance were relatively small. The 75% accuracy condition averaged approximately 62% eyes-off-road time during task performance, and the 60% and 90% averaged approximately 59% eyes-off-road time each.

The 90% accuracy level condition resulted in the lowest number of steering reversals per minute greater than 6 or 12 degrees. A statistical difference was found between all three accuracy level conditions evaluated for the “Steering Reversals > 6°” measure (see Figure 18). However, only the 60% accuracy level differed from both the 75% and 90% accuracy conditions for the “Steering Reversals > 12°” measure (see Figure 20).

Results also showed that recognition accuracy levels could drop to as low as 75% for in-vehicle ASR-operated tasks while maintaining levels of task-function usability and driver preference/acceptance similar to or better than levels recorded for manually-controlled in-vehicle tasks (see Figures 22, 25, 27, and 29). Mean task completion times were shortest for both 90% accuracy conditions (i.e., the 90% accuracy condition that had either substitution errors or rejection errors) and the 75% accuracy condition that had only rejection errors (see Figures 22, and 24). These recognition accuracy conditions resulted in average task completion times that were significantly shorter than identical manually-controlled tasks (see Figure 22). The remaining speech-input conditions (i.e., both 60% accuracy conditions, the 75% accuracy condition that had substitution errors, and the 100% accuracy condition) were statistically similar in task completion times to the manual-input condition. The 75% and 60% accuracy conditions that had substitution errors accounted for all but 11 of the task completion errors recorded in Experiment 2.

The 100% accuracy condition resulted in the longest average task completion times for any of the input conditions tested (see Figure 22). Drivers always experienced this condition on the first day of testing along with the manual-input condition. One possible explanation for this result is that ASR-controlled tasks are unique and drivers were unaccustomed to performing common in-vehicle tasks using speech input. Command vocabulary and syntax may not have been learned well enough the first day to perform speech-input tasks as efficiently as was demonstrated on subsequent days when accuracy levels were lower. The same explanation may hold for manual-input tasks. None of the participants owned a vehicle identical to the test vehicle. Although subjects were trained on manual tasks, manual-input tasks were performed only on the first day of testing. Therefore, task completion times for the 100% accuracy condition and/or the manual-input condition may have been shorter with more practice.

Subjective ratings of usability, comfort, and distraction while performing speech-input tasks were lowest under the 60% recognition condition (see Figures 26, 28, and 30). No statistical difference was found between the 75% and 90% accuracy levels for both ease-of-use ratings and comfort ratings when drivers performed speech-input tasks. However, a statistical difference was found between all three recognition accuracy conditions for subjective ratings of distraction. Again, differences in subjective ratings between recognition accuracy conditions were small. The difference in ratings on the seven-point Likert-type scales between all accuracy conditions and the manual-input condition (i.e., between the 60%, 75%, 90%, 100% recognition accuracy conditions, and manual-input condition) was less than one.

Therefore, the following recommendations are made regarding recognition accuracy for ASR systems used in automotive applications. As a minimum, an automotive ASR system should perform with a 75% recognition accuracy level or better if recognition errors are predominantly rejection errors. If recognition errors are predominantly substitution errors, the system should perform with a 90% recognition accuracy level or better. The preferred level of performance for an automotive ASR system is for recognition accuracy to be 90% or better if both recognition error types are predominant.

Feedback modality. The results from Experiment 1 show that feedback modality had minimal impact on driving performance, and no effect on task-function usability. Visual plus auditory feedback resulted in the lowest number of brake activations per minute, while visual only feedback resulted in the highest number of brake activations per minute, for both the rejection error condition and the substitution error condition. The auditory only feedback condition resulted in a number of brake activations that was between the number of activations for the other two feedback modality conditions across all recognition error types (see Figure 6). The visual component of the visual plus auditory feedback used in Experiment 2 may have been responsible for adversely affecting percent eyes-off-road time (see Figure 15).

The type of feedback modality provided by the ASR system had a minimal effect on driver preference/acceptance of the ASR system tested. Speech-input tasks with visual only feedback resulted in the highest levels of reported distraction across all recognition error types (see Figure 14). The auditory only feedback condition resulted in the lowest levels of reported distraction when either no error occurred or when a substitution error occurred while performing in-vehicle tasks. However, visual plus auditory feedback resulted in the lowest levels of reported distraction when a rejection error occurred during task performance. The difference in subjective ratings was rather small.

Therefore, it is recommended that either a visual plus auditory, or auditory only, feedback method be used in automotive ASR systems. However, if a visual plus auditory feedback is used, the driver should be allowed to select between auditory-only or visual-only feedback so that high levels of user satisfaction and system acceptance can be maintained. Drivers should be allowed to turn off either the visual feedback, the auditory feedback, or both for an automotive ASR system.

Error correction method. The two methods for error correction tested in Experiment 1 had no impact on driving performance, while demonstrating only a modest impact on task-function usability and driver preference/acceptance. Task completion times were shorter on average when substitution errors occurred and drivers used the "Simply Repeat" error correction method (see Figure 8). Task completion times when substitution errors occurred and the "Cancel and Repeat" error correction method was used averaged approximately 32 seconds across all tasks performed. When the "Simply Repeat" error correction method was used under the same substitution error conditions, task completion times averaged approximately 29 seconds across all tasks performed.

The "Simply Repeat" error correction method demonstrated significantly higher subjective ratings of usability than the "Cancel and Repeat" error correction method (see Figure 10). However, the difference in subjective ratings between error correction methods was again rather small. The difference in ratings between error correction methods on the five-point Likert-type scale for ease-of-use was less than 0.50.

Therefore, it is recommended that most speech command inputs should be correctable through command repetition (i.e., a "Simply Repeat" type error correction method). However, for some types of tasks, this may be unfeasible (e.g., entering a telephone number using a hands-free cellular phone, or for tasks that are safety critical) and a method for canceling or deleting prior command inputs should be allowed.

Input condition. The type of input condition, either speech or manual, had a significant effect on driver eye-movement behavior. Percent eyes-off-road time during task performance was significantly greater for speech-input tasks than for manual-input tasks (see Figure 15). No statistical difference was found between any of the seven speech-input conditions. Mean percent eyes-off-road for manual tasks was approximately 36%, while mean percent eyes-off-road averaged across all speech-input conditions was approximately 58%. These results contradict the hypothesis that speech-input tasks would allow drivers to maintain higher percent eyes-<u>on</u>-road time compared with similar manual-input tasks during performance of an in-vehicle task.

The novelty of speech input for drivers may have caused them to look inside the vehicle for longer durations to confirm that a task was performed correctly or to observe what happened when they spoke commands. A similar effect on percent eyes-off-road was found for another novel automotive device, the head-up display (Kiefer, 1991). However, the effect went away after limited experience with the novel device (after approximately 35 minutes of driving with a HUD). The novelty effect of automobiles with ASR controls may be more robust than that of HUD-equipped vehicles. Further study is required to determine if percent eyes-off-road changes for speech-input tasks as drivers become more familiar with and more comfortable operating ASR-controlled in-vehicle tasks.

Another explanation for why mean percent eyes-off-road time was greater for speech-input tasks may have something to do with the type of feedback modality provided by the ASR system. A visual plus auditory feedback method was used in Experiment 2. The visual component of the feedback may have compelled drivers to look inside the vehicle during task performance, resulting in the greater percent eyes-off-road time for speech-input tasks recorded. Further research would be required to determine if an auditory-only type of feedback could reduce eyes-off-road time compared to a visual plus auditory feedback method.

It is important to note that although percent eyes-off-road was greater for speech-input tasks, no large effects on driving performance were found. This supports the hypothesis that this difference in eyes-off-road time was largely due to novelty effects, and that the visual attention used in performing the ASR tasks was drawn from "spare" resource capacity.

In general, the majority of speech-input tasks tested in Experiment 2 showed no significant differences in driving performance, task-function usability, and driver preference/acceptance when compared with identical manual-input tasks. In some instances, however, certain speech-input conditions (e.g., the 90% accuracy condition that had only rejection errors, or the 100% accuracy condition) resulted in significant improvements over manual-input tasks for several of the dependent measures recorded (see Figures 17, 22, 27, and 29). These improvements, although statistically significant, were again relatively small and may be of little practical significance.

Driver age. Age is an important variant in driving performance. The effects of aging are well documented in the literature. Sensory, perceptual, cognitive, and psychomotor impairments that affect driving occur more frequently within the older driving population. Age was found to significantly affect measures of driving performance and task-function usability recorded in Experiment 2, but age had no effect on measures of driver preference/acceptance.

Measures of driving performance affected most by age were steering-input behavior and speed-maintenance behavior. Steering variability was significantly greater for younger drivers (ages 21 to 27 years) than older drivers (ages 65 to 78) while performing in-vehicle tasks. Steering variability during task performance for younger drivers was 0.75 degrees, while steering variability during task performance for older drivers was 0.51 degrees. Number of steering reversals per minute greater than 6 and 12 degrees was greater for younger drivers compared to the older drivers.

Younger drivers demonstrated greater accelerator position variability (younger drivers averaged an accelerator position variance of 0.017, and older drivers averaged an accelerator position variance of 0.013 for accelerator positions normalized between zero and one). In addition, older drivers demonstrated significantly more brake activations per minute than younger drivers when performing in-vehicle tasks (2.43 brake activations per minute versus 1.10 brake activations, respectively).

Although driver age had a significant effect on driving performance, the differences in measures between younger drivers and older drivers were small enough to be of little practical significance. The age effects found in Experiment 2 were predicted and have been discovered by numerous studies on aging and driving safety.

Interestingly, the only dependent variable affected by age that was related to ASR systems was task completion errors. Older drivers accounted for the significant proportion of task completion errors recorded (79 errors out of a total 111 recorded errors). Therefore, driver age is still an important variable to consider for issues of task-function usability in the design of ASR systems for automotive applications.

Decision trees. The decision-tree tool failed to provide a satisfactory level of prediction accuracy for task-function input-modality allocation. The large majority of task conditions failed to show any statistical differences between manual-input and the speech-input conditions for each of the dependent measures, regardless of the overall decision tree recommendation. The tasks for which a difference was found between manual input and the speech input conditions showed that overall the decision trees recommended the appropriate input modality a total of 15 times, and the decision trees recommended the opposite input modality (compared with the empirical results) a total of 9 times across all of the dependent measures recorded. In other words, when a difference was found between manual and speech input for a given task, the decision trees recommended the appropriate input modality 62.5 percent of the time. However, it should be noted that a difference between manual input and the speech input condition was discovered for only 25 task conditions out of a total 128 conditions recorded across all dependent measures.

The decision trees were developed and used to make recommendations for task-function input modality prior to any data collection in the test vehicle. Development of the decision trees was a first cut attempt by the author to provide designer with a usable tool for the development of optimal interfaces for in-vehicle systems. The decision points for the trees were based on information from the literature and expert opinion. The next logical step is to refine the decision trees and/or methodology to improve the predictive power of the tool, and to provide designers of in-vehicle controls with an efficient and usable method for determining task-function input-modality allocation.

Human factors recommendations. Table 14 presents the minimal and preferred ASR system specifications for each system parameter that was investigated in the empirical research. The table summarizes the design recommendations discussed above and formats them in a useful quick-reference tool for designers and engineers of speech-input in-vehicle tasks. The recommendations are based on the results from Experiments 1 and 2.

Table 14. Minimal and preferred design parameter specifications for ASR systems considered for use in automobiles.

ASR System Design Parameter	ASR System Specifications: Minimal	ASR System Specifications: Preferred
Feedback Modality	The system feedback should have an auditory component to ensure driver awareness of command recognition.	The system feedback should have both an auditory and visual component and allow the driver to select between either, or both.
Recognition Accuracy	The system may perform with a 75% or higher recognition accuracy if substitution errors are uncommon (i.e., only rejection errors are common), and drivers can control system feedback modality (i.e., select modality of feedback or turn off feedback entirely). Substitution error rate should be less than 10%.	The system should perform with a 90% or higher recognition accuracy if recognition error type can not be controlled.
Recognition Errors	If the system configuration is such that rejection errors occur, the system should perform with a 75% or greater recognition accuracy. If the system configuration is such that substitution errors occur, the system should perform with a 90% or greater recognition accuracy.	The system should minimize the number of rejection and substitution errors allowed.
Error Correction Method	The system should allow drivers to "cancel" or "delete" commands that result in substitution errors.	The system should allow drivers to simply repeat the command within a specific time period to correct substitution errors when they occur, and to "cancel" or "delete" commands that result in substitution errors that are safety-critical in nature.

CONCLUSIONS

This research has important implications for the use of automatic speech recognition technology in automotive applications. Knowledge gained from the two on-road studies can be applied to the design of in-vehicle systems that are controlled using ASR technology. The decision trees, developed as part of the research effort, fall short of providing designers and engineers of automotive secondary systems with a reliable tool for initially determining task-function input-modality allocation. However, the decision tool approach should not be abandoned, and further research is warranted to refine such a tool that aids in the design and development of in-vehicle systems. Results from this research should make it possible to improve the design of automotive secondary-control systems. Improvements in the design of automotive secondary-control systems can have a significant impact on driving safety, system usability, and customer satisfaction. This research also attempted to provide evidence that supports (1) the claim that speech-controlled in-vehicle tasks may improve driver safety by minimizing the visual (drivers' eyes-off-road time) and the manual demands of performing current in-vehicle tasks and, most important, (2) the claim that speech systems can operate with significantly less than 100% recognition accuracy, and still maintain levels of driving performance, system usability, and driver acceptance comparable to or better than current manual-input tasks.

The research effort was a logical progression in the assessment of attentional demands (i.e., visual, manual, and cognitive demands) required to perform various in-vehicle tasks. Similar research has been performed that evaluated attentional demands of manual-controlled in-vehicle tasks (see Wierwille, 1993). Wierwille and his colleagues conducted a series of on-road and simulator experiments to evaluate the effects that performing manual in-vehicle tasks have on driving performance and attentional demand. The current research addressed how attentional demands are affected by performing in-vehicle tasks using speech input, as well as current manual-control methods.

Manipulation of various ASR system parameters was shown to significantly affect measures of driving performance. However, from a practical viewpoint, ASR system parameters had a nominal effect on driving performance. Differences measured in driving performance brought on by changes in ASR system parameters were small enough that alternative ASR system designs can be considered without impacting driving performance. The speech-input methods tested were not practically different from current manual-input methods used in performing identical in-vehicle tasks for any of the measures of driving performance recorded. Therefore, no benefits for ASR systems improving driving safety/performance compared to manual-control systems can be claimed.

Manipulation of the various ASR system parameters was shown to significantly affect measures of task-function usability. Automatic speech recognition system parameters had a significant impact on the usability of the in-vehicle tasks performed. Criteria such as task-completion times and task completion errors were shown to be different between speech-input and manual-input control methods. Therefore, trade-offs in ASR system designs versus manual-control system designs can be evaluated in terms of usability, and the system design that results in the highest levels of usability should be selected for use in the automotive environment.

Finally, manipulation of the various ASR system parameters was shown to significantly affect measures of driver preference/acceptance. However, the differences between subjective opinion data were relatively small and may not be of any practical significance. Further research is warranted to determine if the long-term use of ASR systems with less than optimal design parameters would result in significantly lower preference/acceptance data compared to the data collected in this research effort. If system performance does not improve with use (e.g., through the use of speaker-adaptive ASR systems or user-enrolled ASR systems), it is hypothesized that ASR systems that continually perform at 75% recognition accuracy with predominantly rejection errors would frustrate drivers over time, significantly lowering their subjective ratings of the system.

Future Research Applications

Future research could proceed in several directions. The methodology used in this research might be refined and used to evaluate future speech recognition interfaces where a variety of in-vehicle tasks are performed and system parameters of interest are manipulated. Guidelines for the design and implementation of system variants could be developed from this research. The "wizard of oz" technique employed in this research is a useful methodology for evaluating system designs before they are fully operational. This methodology can be applied to a number of other human-machine systems where design parameters require empirical testing so that optimized design decision-making results.

A second research approach should be to evaluate the long-term effects of using ASR systems in automotive applications. Speech systems that demonstrate lower levels of performance could be evaluated in longitudinal studies to determine how measures of driving performance, task-function usability, and driver preference/acceptance change as drivers interact with the system under more realistic conditions. This would be a logical step in the progression of research on ASR systems used for automotive applications to determine system performance parameters that are acceptable to drivers, and that do not negatively impact driving safety.

Finally, the decision trees that supported task-function input-modality allocation for the current research effort should be refined. The decision trees need to be made more reliable for recommending a given in-vehicle task to be controlled using speech input instead of manual input, or vice versa. A highly reliable and validated decision tool will help improve the overall design process for in-vehicle systems, and could potentially shorten the time taken to design optimally controlled in-vehicle systems.

REFERENCES

Allen, R.W., Stein, A.C., Rosenthal, T.J., Ziedman, D., Torres, J.F., and Halati, A. (1991). A human factors simulation investigation of driver route diversion and alternate route selection using in-vehicle navigation systems. In Proceedings of the Vehicle Navigation and Information Systems Conference (VNIS '91) (SAE paper 912731) (pp. 9-26). Warrendale, PA: Society of Automotive Engineers.

Aretz, A.J. (1983). A comparison of manual and vocal response modes for the control of aircraft systems. In Proceedings of the Human Factors Society 27th Annual Meeting (pp. 97-101). Santa Monica, CA: Human Factors Society.

Atal, B.S. (1994). Speech technology in 2001. In D.B. Roe and J.G. Wilpon (Eds.), Voice Communication Between Humans and Machines (pp. 467-481). Washington, D.C.: National Academy Press.

Baber, C. (1991). Speech Technology in Control Room Systems: A Human Factors Perspective. New York: Ellis Horwood.

Baber, C., Usher, D.M., Stammers, R.B., and Taylor, R.G. (1992). Feedback requirements for automatic speech recognition in the process control room. International Journal of Man-Machine Studies, 37, 703-719.

Bailey, R.W. (1989). Human performance engineering: Using human factors/ergonomics to achieve computer system usability (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Beek, B., Neuberg, E.P., and Hodge, D.C. (1977). An assessment of the technology of automatic speech recognition for military applications. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-25(4), 310-322.

Bekker, M.M., Van Nes, F.L., and Juola, J.F. (1995). A comparison of mouse and speech input control of a text-annotation system. Behaviour and Information Technology, 14, 14-22.

Bennett, R.W., Greenspan, S.L., Syrdal, A.K., Tschirgi, J.E., and Wisowaty, J.J. (1989). Speaking to, from, and through computers: Speech technologies and user-interface design. AT&T Technical Journal, 68(5), 17-30.

Bernold, L.E. (1993). Speech-based data-entry systems for construction. Journal of Computing in Civil Engineering, 7, 404-419.

Biermann, A.W., Fineman, L., and Heidlage, J.F. (1992). A voice- and touch-driven natural language editor and its performance. International Journal of Man-Machine Studies, 37, 1-21.

Brandt, D.J. (1987). Voice controlled instrument cluster display. Unpublished bachelor thesis, GMI Engineering and Management Institute, Flint, MI.

Brems, D.J., Rabin, M.D., and Waggett, J.L. (1995). Using natural language conventions in the user interface design of automatic speech recognition systems. Human Factors, 37, 265-282.

Bristow, G. (1986). The speech recognition problem. In G. Bristow (Ed.), Electronic Speech Recognition -- Techniques, Technology, and Applications (pp. 3-17). London: Collins.

Brookhuis, K.A., de Vries, G., and de Waard, D. (1991). The effects of mobile telephoning on driving performance. Accident Analysis and Prevention, 23, 309-316.

Brown, I.D., Tickner, A.H., and Simmonds, D.C.V. (1969). Interference between concurrent tasks of driving and telephoning. Journal of Applied Psychology, 53, 419-424.

Casali, S.P., Williges, B.H., and Dryden, R.D. (1990). Effects of recognition accuracy and vocabulary size of a speech recognition system on task performance and user acceptance. Human Factors, 32, 183-196.

Cochran, D.J., Riley, M.W., and Stewart, L.A. (1980). An evaluation of the strengths, weaknesses, and uses of voice input devices. In Proceedings of the Human Factors Society 24th Annual Meeting (pp. 190-194). Santa Monica, CA: Human Factors Society.

Cohen, A., and Mandel, A.F. (1982). Considerations for integrating speech I/O technology into factory automation systems. In Wescon/82 Conference Record (pp. 29-1/1-8). El Segundo, CA: Electron.

Cohen, P.R., and Oviatt, S.L. (1994). The role of voice in human-machine communication. In D.B. Roe and J.G. Wilpon (Eds.), Voice Communication Between Humans and Machines (pp. 34-75). Washington, D.C.: National Academy Press.

Cosky, M.J., Lively, B.L., Roberts, L.A., and Wattenbarger, B.L. (1995). Talking to machines today and tomorrow: Designing for the user. AT&T Technical Journal, 74(2), 81-91.

Costet, J.R. (1982). Voice techniques on board aircraft: An experimental approach. In Proceedings of the IEEE 1982 National Aerospace and Electronics Conference (pp. 339-342). New York: IEEE.

Damos, D. (1985). The effect of asymmetric transfer and speech technology on dual-task performance. Human Factors, 27, 409-421.

Damos, D. (1986). The effect of using voice generation and recognition systems on the performance of dual tasks. Ergonomics, 29, 1359-1370.

Dingus, T.A. (1995). A meta-analysis of driver eye-scanning behavior while navigating. In Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting (pp. 1127-1131). Santa Monica, CA: Human Factors Society.

Dingus, T.A., Antin, J.F., Hulse, M.C., and Wierwille, W.W. (1989). Attentional demand requirements of an automobile moving-map navigation system. Transportation Research - A, 23A, 301-315.

Dingus, T.A., Gellatly, A.W., and Reinach, S. (1997). Human-computer interaction applications for intelligent transportation systems. In M. Helander (Ed.), Handbook of Human-Computer Interaction (2nd. Ed.).

Dingus, T.A., Hulse, M.C., Krage, M.K., Szczublewski, F.E., and Berry, P. (1991). A usability evaluation of navigation and information system "pre-drive" functions. In Proceedings of the Vehicle Navigation and Information Systems Conference (VNIS '91) (SAE paper 912794) (pp. 527-536). Warrendale, PA: Society of Automotive Engineers.

Dingus, T.A., Hulse, M.C., McGehee, D.V., and Manakkal, R. (1994). Driver performance results from the TravTek IVHS camera car evaluation study. In Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting (pp. 1118-1122). Santa Monica, CA: Human Factors Society.

Doddington, G. (1980). Whither speech recognition? In W.A. Lea (Ed.), Trends in Speech Recognition (pp. 556-561). Englewood Cliffs, NJ: Prentice-Hall.

Doddington, G.R., and Schalk, T.B. (1981). Speech recognition: Turning theory to practice. IEEE Spectrum, 18(9), 26-32.

Ericsson, K.A., and Simon, H.A. (1984). Protocol Analysis: Verbal Reports as Data. Cambridge, MA: MIT Press.

Fairclough, S.H., Ashby, M.C., and Parkes, A.M. (1993). In-vehicle displays, visual workload and usability evaluation. In A.G. Gale (Ed.), Vision in Vehicles IV (pp. 245-254). Amsterdam, Netherlands: Elsevier.

Fink, D.F. (1981). Automatic speech recognition systems -- Evaluation and application. In Wescon/81 Conference Record (pp. 6/2 1-8). El Segundo, CA: Electron.

Forren, M.G., and Mitchell, C.M. (1986). Multi-modal interfaces in supervisory control. In Proceedings of the Human Factors Society 30th Annual Meeting (pp. 317-321). Santa Monica, CA: Human Factors Society.

Furui, S. (1994). Toward the ultimate synthesis/recognition system. In D.B. Roe and J.G. Wilpon (Eds.), Voice Communication Between Humans and Machines (pp. 450-466). Washington, D.C.: National Academy Press.

Gould, J.D., Conti, J., and Hovanyecz, T. (1983). Composing letters with a simulated listening typewriter. Communications of the ACM, 26, 295-308.

Haller, R., Bouis, D., and Heintz, F. (1981). More driving safety with a better designed trip computer keyboard. In Proceedings of the Human Factors Society 25th Annual Meeting (pp. 721-725). Santa Monica, CA: Human Factors Society.

Hauptman, A.G., and McAvinney, P. (1993). Gestures with speech for graphic manipulation. International Journal of Man-Machine Studies, 38, 231-249.

Howell, D.C. (1992). Statistical Methods for Psychology (3rd ed.). Belmont, CA: Duxbury Press.

Isenberg, D., Yntena, D., and Wiesen, R. (1984). Designing speech recognition interfaces for talkers and tasks. In G. Salvendy (Ed.), Human-Computer Interaction (pp. 455-458). Amsterdam: Elsevier.

Jones, D.M., Frankish, C.R., and Hapeshi, K. (1992). Automatic speech recognition in practice. Behaviour and Information Technology, 11, 109-122.

Jones, D., Hapeshi, K., and Frankish, C. (1989). Design guidelines for speech recognition interfaces. Applied Ergonomics, 20, 47-52.

Juang, B.H., Perdue, R.J., and Thomson, D.L. (1995). Deployable automatic speech recognition systems: Advances and challenges. AT&T Technical Journal, 74(2), 45-55.

Kames, A.J. (1979). A study of the effects of mobile telephone use and control unit design on driving performance. IEEE Transactions on Vehicular Technology, VT-27(4), 282-287.

Kamm, C. (1994). User interfaces for voice applications. In D.B. Roe and J.G. Wilpon (Eds.), Voice Communication Between Humans and Machines (pp. 422-442). Washington, D.C.: National Academy Press.

Keppel, G. (1991). Design and Analysis: A researcher's handbook (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Kiefer, R.J. (1991). Effect of a head-up versus head-down digital speedometer on visual sampling behavior and speed control performance during daytime automobile driving (SAE Technical Paper Series No. 910111). Warrendale, PA: Society of Automotive Engineers.

Kiefer, R.J., and Gellatly, A.W. (1996). Toward quantifying the real-world consequences of the claimed "eyes-on-road" benefit of head-up displays (SAE Technical Paper Series No. 960946). Warrendale, PA: Society of Automotive Engineers.

Kiger, S.M., Rockwell, T.H., and Tijerina, L. (1995). Developing baseline data on heavy vehicle driver visual workload. In Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting (pp. 1112-1116). Santa Monica, CA: Human Factors Society.

Kirwan, B., and Ainsworth, L.K. (Eds.). (1992). A Guide to Task Analysis. London: Taylor and Francis.

Kleinmuntz, D.N. (1987). Human decision processes: Heuristics and task structure. In P.A. Hancock (Ed.), Human Factors Psychology (pp. 123-157). Amsterdam, Netherlands: Elsevier.

Kobe, G. (1995). Buick's electric toybox. Automotive Industries, 175(6), 78-80.

Kurokawa, K. (1990). Development of an evaluation program for automotive instrument panel design. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Lea, W.A. (1980a). Speech recognition: Past, present, and future. In W.A. Lea (Ed.), Trends in Speech Recognition (pp. 39-98). Englewood Cliffs, NJ: Prentice-Hall.

Lea, W.A. (1980b). The value of speech recognition systems. In W.A. Lea (Ed.), Trends in Speech Recognition (pp. 3-18). Englewood Cliffs, NJ: Prentice-Hall.

Lee, K.F., Hauptman, A.G., and Rudnicky, A.I. (1990). The spoken word: Replace the "look and feel" of GUIs with the "ask and tell" of voice interfaces. BYTE, 15(7), 225-232.

Makhoul, J., and Schwartz, R. (1994). State of the art in continuous speech recognition. In D.B. Roe and J.G. Wilpon (Eds.), Voice Communication Between Humans and Machines (pp. 165-198). Washington, D.C.: National Academy Press.

Marin-Lamellet, C., and Dejeammes, M. (1995). The processing of complex guidance symbols by elderly drivers: A simulator-based study and an evaluation of the CARMINAT guidance system by the European Community DRIVE-EDDIT project. In Proceedings of the Vehicle Navigation and Information Systems Conference (VNIS '95) (pp. 110-117). Piscataway, NJ: IEEE.

Markowitz, J.A. (1995). Talking. Journal of Systems Management, 46(3), 9-13.

Martin, G.L. (1989). The utility of speech input in user-computer interfaces. International Journal of Man-Machine Studies, 30, 355-375.

Martin, J., Long, J., and Broome, D. (1984). The division of attention between a primary tracking task and secondary tasks of pointing with a stylus or speaking in a simulated ship's-gunfire-control task. Ergonomics, 27, 397-408.

Martin, T.B. (1976). Practical applications of voice input to machines. Proceedings of the IEEE, 64, 487-501.

Martin, T.B., and Welch, J.R. (1980). Practical speech recognizers and some performance effectiveness parameters. In W.A. Lea (Ed.), Trends in Speech Recognition (pp. 24-38). Englewood Cliffs, NJ: Prentice-Hall.

Mollenhauer, M.A., Lee, J., Cho, K., Hulse, M.C., and Dingus, T.A. (1994). The effects of sensory modality and information priority on in-vehicle signing and information systems. In Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting (pp. 1072-1076). Santa Monica, CA: Human Factors Society.

Mountford, S.J., and North, R.A. (1980). Voice entry for reducing pilot workload. In Proceedings of the Human Factors Society 24th Annual Meeting (pp. 185-189). Santa Monica, CA: Human Factors Society.

Mountford, S.J., North, R.A., Metz, S.V., and Warner, N. (1982). Methodology for exploring voice-interactive avionic tasks: Optimizing interactive dialogues. In Proceedings of the Human Factors Society 26th Annual Meeting (pp. 207-211). Santa Monica, CA: Human Factors Society.

Mountford, S.J., Schwartz, J., and Graffunder, K. (1983). Evaluation of speech technology for automatic target recognition. In Proceedings of the Human Factors Society 27th Annual Meeting (pp. 965-969). Santa Monica, CA: Human Factors Society.

Mourant, R.R., Herman, M., and Moussa-Hamouda, E. (1980). Direct looks and control location in automobiles. Human Factors, 22, 417-425.

Murray, J.T., and Van Praag, J., and Gilfoil, D. (1983). Voice versus keyboard control of cursor motion. In Proceedings of the Human Factors Society 27th Annual Meeting (pp. 103). Santa Monica, CA: Human Factors Society.

Newell, J.L., Arnott, J.L., Carter, K., and Cruickshank, G. (1990). Listening typewriter simulation studies. International Journal of Man-Machine Studies, 33, 1-19.

Ng, L., and Barfield, W. (1993). User information requirements for intelligent vehicle highway systems as a function of driver category. In Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting (pp. 1077-1081). Santa Monica, CA: Human Factors Society.

Paelke, G.M. (1993). Comparison of route guidance destination entry methods. In Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting (pp. 569-573). Santa Monica, CA: Human Factors Society.

Pallett, D. (1986). Assessing the performance of recognisers. In G. Bristow (Ed.), Electronic Speech Recognition -- Techniques, Technology, and Applications (pp. 277-306). London: Collins.

Peckham, J.B. (1984). Automatic speech recognition -- A solution in search of a problem? Behavior and Information Technology, 3, 145-152.

Peckham, J. (1986). Human factors in speech recognition. In G. Bristow (Ed.), Electronic Speech Recognition -- Techniques, Technology, and Applications (pp. 172-187). London: Collins.

Pelton, G.E. (1993). Voice Processing. New York: McGraw-Hill.

Perez, W.A., and Mast, T.M. (1992). Human factors and advanced traveler information systems (ATIS). In Proceedings of the Human Factors Society 36th Annual Meeting (pp. 1073-1077). Santa Monica, CA: Human Factors Society.

Poock, G.K. (1986). Speech recognition research, applications, and international efforts. In Proceedings of the Human Factors Society 30th Annual Meeting (pp. 1278-1283). Santa Monica, CA: Human Factors Society.

Poulton, E.C. (1982). Influential companions: Effects of one strategy on another in the within-subjects designs of cognitive psychology. Psychological Bulletin, 91, 673-690.

Quinnell, R.A. (1995). Speech recognition: No longer a dream but still a challenge. EDN, 40(2), 41-46.

Rabiner, L.R. (1995). Toward vision 2001: Voice and audio processing considerations. AT&T Technical Journal, 74(2), 4-13.

Roe, D.B., and Wilpon, J.G. (1993). Whither speech recognition: The next 25 years. IEEE Communications Magazine, 31(11), 54-62.

Sanders, M.S., and McCormick, E.J. (1987). Human Factors in Engineering and Design (6th ed.). New York: McGraw-Hill.

SAS Institute Inc. (1995). SAS/STAT User’s Guide, Release 6.03 Edition. Cary, NC: SAS Institute Inc.

Schafer, R.W. (1994). Scientific bases of human-machine communication by voice. In D.B. Roe and J.G. Wilpon (Eds.), Voice Communication Between Humans and Machines (pp. 15-33). Washington, D.C.: National Academy Press.

Schurick, J.M., Maynard, J.F., and Williges, B.H. (1983). Feedback techniques for voice input in computer-based systems. In Proceedings of the Human Factors Society 27th Annual Meeting (pp. 102). Santa Monica, CA: Human Factors Society.

Schurick, J.M., Williges, B.H., and Maynard, J.F. (1985). User feedback requirements with automatic speech recognition. Ergonomics, 28(11), 1543-1555.

Serafin, C., Wen, C., Paelke, G., and Green, P. (1993). Car phone usability: A human factors laboratory test. In Proceedings of the Human Factors Society 37th Annual Meeting (pp. 220-224). Santa Monica, CA: Human Factors Society.

Siegel, S., and Castellan, N.J. (1988). Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill.

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

Simpson, C.A., McCauley, M.E., Roland, E.F., Ruth, J.C., and Williges, B.H. (1987). Speech controls and displays. In G. Salvendy (Ed.), Handbook of Human Factors (pp. 1490-1525). New York: Wiley.

Sivak, M., Flannagan, M., and Gellatly, A.W. (1991). Reaction times to high-contrast brake lamps (SAE Technical Paper Series No. 910821). In D. Hoffmeister (Ed.), Vehicle Lighting Design for Optimal Visibility and Performance: SP-857 (pp. 19-23). Warrendale, PA: Society of Automotive Engineers.

Snyder, H.L., and Monty, R.W. (1986). A methodology for road evaluation of automobile displays. In A.G. Gale (Ed.), Vision in Vehicles (pp. 227-237). Amsterdam, Netherlands: Elsevier.

Spine, T.M., Maynard, J.F., and Williges, B.H. (1983). Error correction strategies for voice recognition. In Proceedings of the Voice Data Entry Systems Application Conference. Chicago, IL: American Voice I/O Society.

Spine, T.M., Williges, B.H., and Maynard, J.F. (1984). An economical approach to modeling speech recognition accuracy. International Journal of Man-Machine Studies, 21, 191-202.

Steiner, B.A., Burkhart, B.J., and Berson, B.L. (1983). Voice feedback protocols. In Proceedings of the Human Factors Society 27th Annual Meeting (pp. 108-111). Santa Monica, CA: Human Factors Society.

Tabachnick, B.G., and Fidell, L.S. (1989). Using Multivariate Statistics (2nd ed.). New York: Harper and Row.

Tijerina, L., Kiger, S.M., Rockwell, T.H., and Tornow, C. (1995). Workload assessment of in-cab text message system and cellular phone use by heavy vehicle drivers on the road. In Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting (pp. 1117-1121). Santa Monica, CA: Human Factors Society.

Tsang, P.S., Hartzell, E.J., and Rothschild, R.A. (1985). To speak or not to speak: A multiple resource perspective. In Proceedings of the Human Factors Society 29th Annual Meeting (pp. 76-80). Santa Monica, CA: Human Factors Society.

Vail, R.E. (1986). Suitability of voice-activated hand controls for the physically handicapped driver. In Proceedings of the Human Factors Society 30th Annual Meeting (pp. 662-666). Santa Monica, CA: Human Factors Society.

Vidulich, M.A. (1988). Speech responses and dual-task performance: Better time-sharing or asymmetric transfer? Human Factors, 30, 517-529.

Walker, J., Alicandri, E., Sedney, C., and Roberts, K. (1991). In-vehicle navigation devices: Effects on the safety of driver performance. In Proceedings of the Vehicle Navigation and Information Systems Conference (VNIS '91) (SAE paper 912793) (pp. 499-525). Warrendale, PA: Society of Automotive Engineers.

Wattenbarger, B.L., Garberg, R.B., Halpern, E.S., and Lively, B.L. (1993). Serving customers with automatic speech recognition -- Human factors issues. AT&T Technical Journal, 72(3), 28-41.

Wharton, C., Marics, M., and Engelbeck, G. (1993). Speech recognition vocabulary scoping for automated call routing. In Proceedings of the Human Factors Society 37th Annual Meeting (pp. 240-243). Santa Monica, CA: Human Factors Society.

Wickens, C.D. (1992). Engineering Psychology and Human Performance (2nd ed.). New York: HarperCollins.

Wickens, C.D., Sandry, D.L., and Vidulich, M. (1983). Compatibility and resource competition between modalities of input, central processing, and output. Human Factors, 25, 227-248.
Wickens, C.D., Vidulich, M., and Sandry, D. (1981). Factors influencing the performance advantage of speech technology. In Proceedings of the Human Factors Society 25th Annual Meeting (pp. 705-709). Santa Monica, CA: Human Factors Society.
Wickens, C.D., Vidulich, M., and Sandry-Garza, D. (1984). Principles of S-C-R compatibility with spatial and verbal tasks: The role of display-control location and voice-interactive display-control interfacing. Human Factors, 26, 533-543.
Wierwille, W.W. (1993). Visual and manual demands of in-car controls and displays. In B. Peacock and V. Karwowski (Eds.), Automotive Ergonomics (pp. 299-320). London: Taylor and Francis.
Williges, R.C. (1981). Development and use of research methodologies for complex system/simulation experimentation. In J. Moraal and K.F. Kraiss (Eds.), Manned Systems Design: Methods, Equipment, and Applications (pp. 59-87). New York: Plenum.
Williges, R.C., and Williges, B.H. (1989). Integrated research paradigm for complex experimentation. In Proceedings of the Human Factors Society 33rd Annual Meeting (pp. 606-609). Santa Monica, CA: Human Factors Society.
Zoltan-Ford, E. (1984). Reducing variability in natural-language interactions with computers. In Proceedings of the Human Factors Society 28th Annual Meeting (pp. 768-772). Santa Monica, CA: Human Factors Society.
Zwahlen, H.T., and DeBald, D.P. (1986). Safety aspects of sophisticated in-vehicle information displays and controls. In Proceedings of the Human Factors Society 30th Annual Meeting (pp. 256-260). Santa Monica, CA: Human Factors Society.

APPENDICES

APPENDIX A: Decision Trees

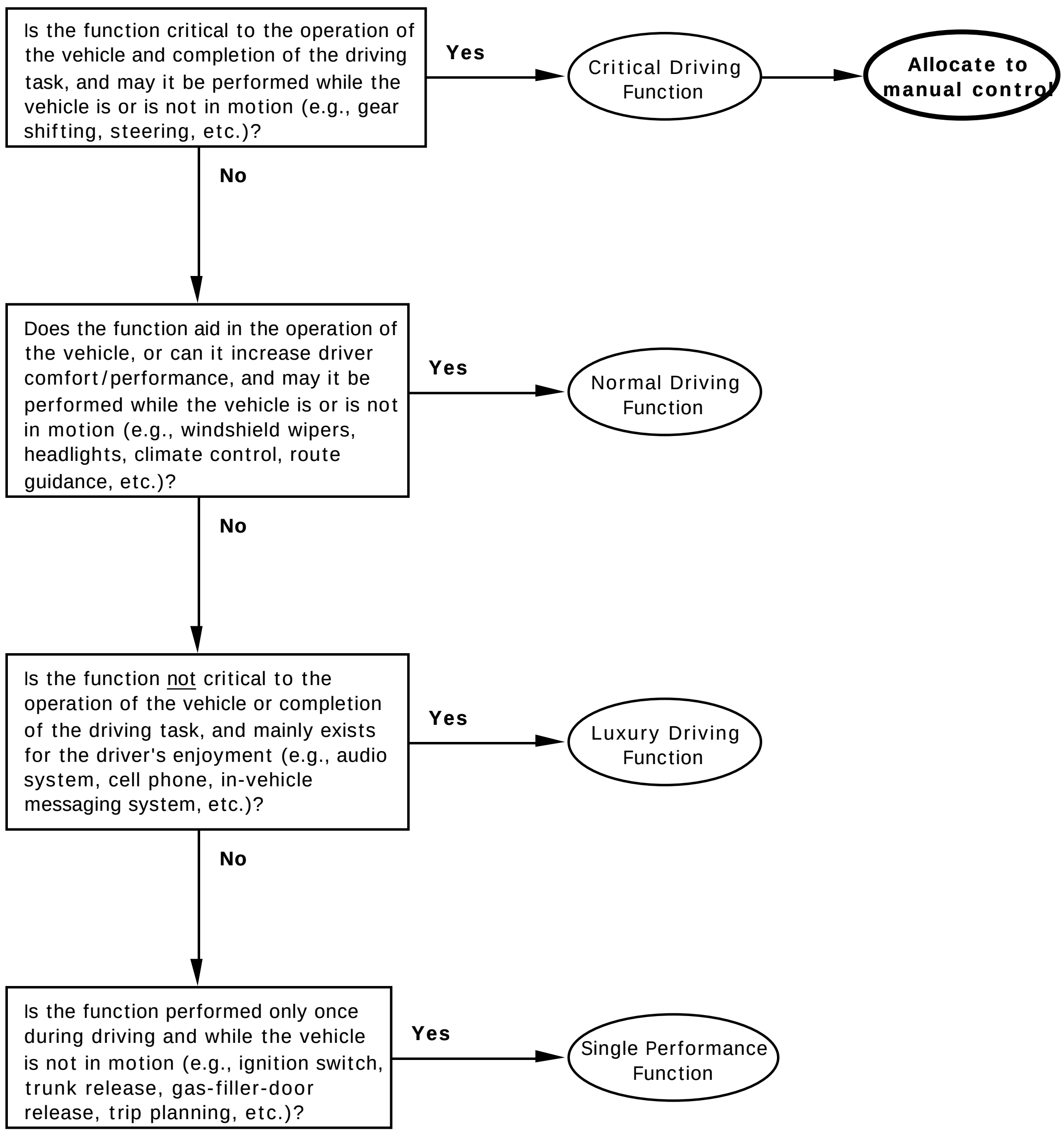

Figure A-1. Driving functions decision tree.

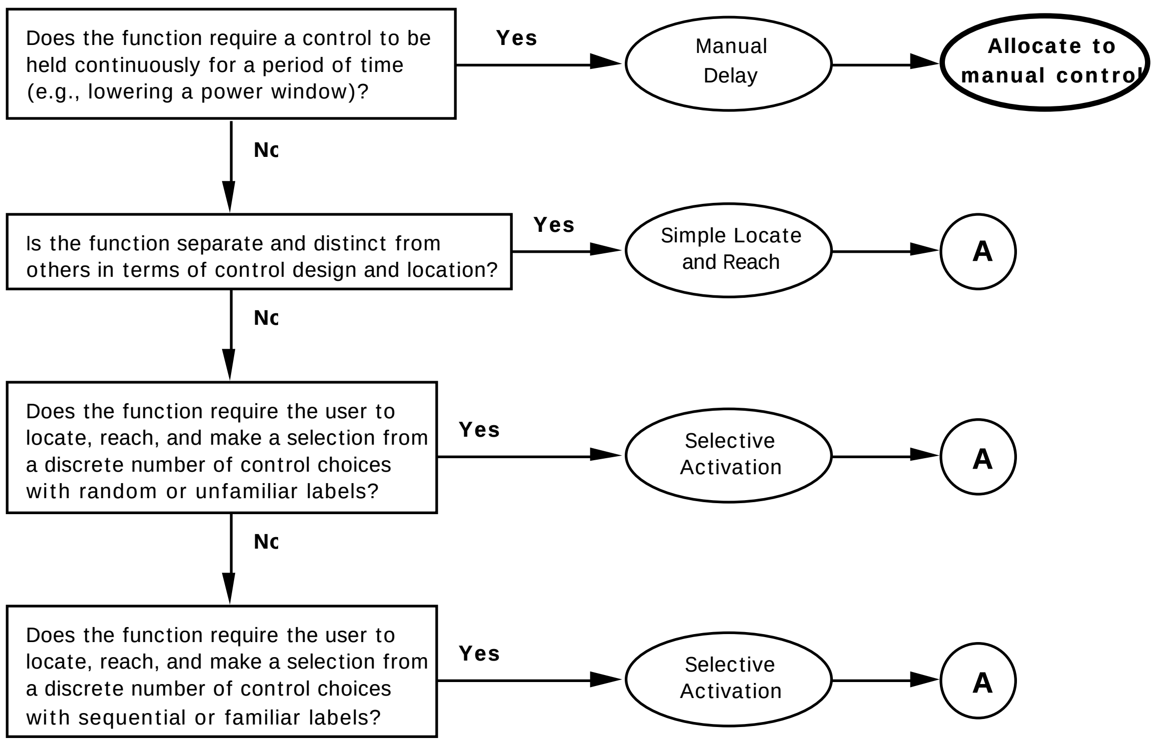

Figure A-2. Behavioral categorization and decision allocation decision tree.

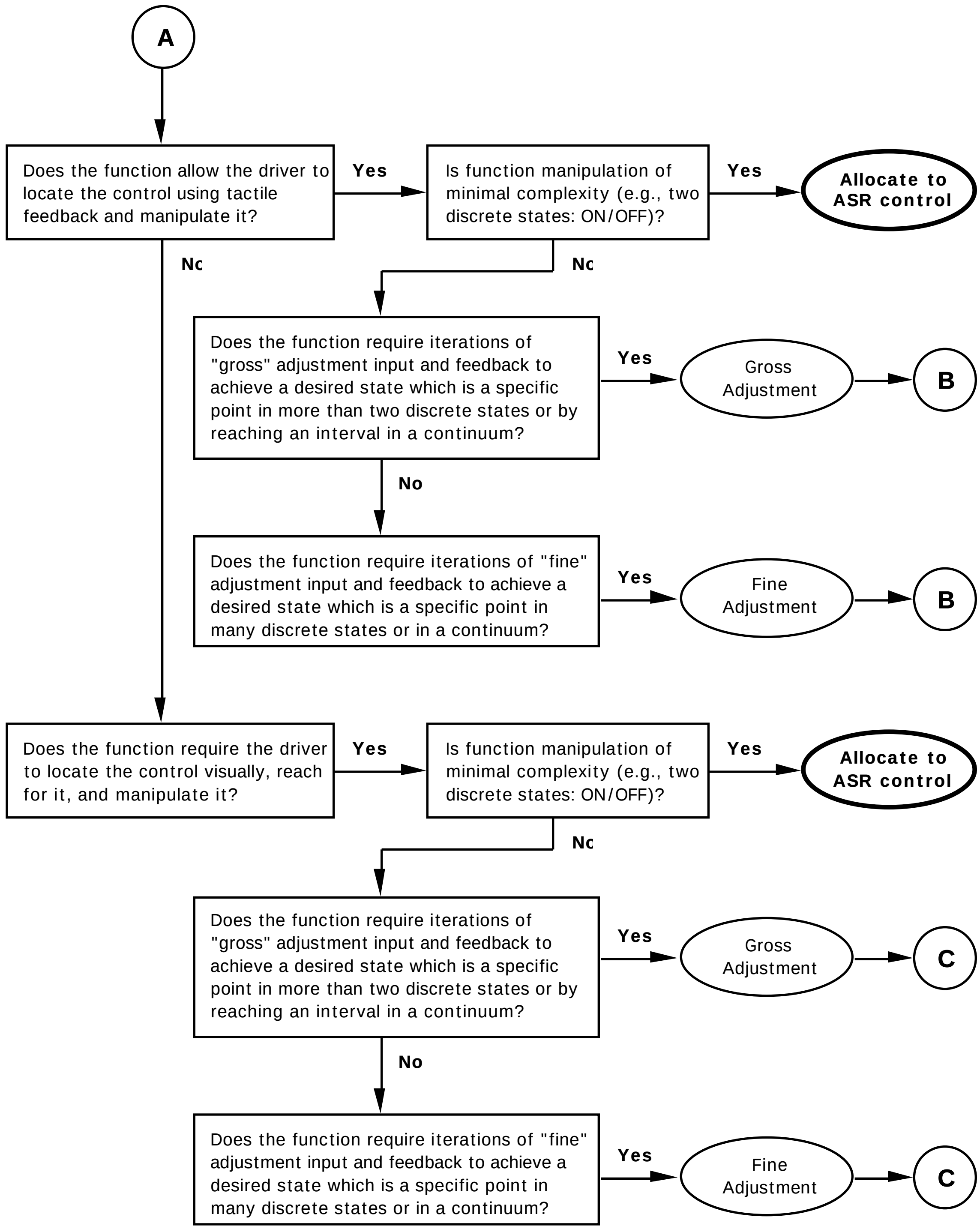

Figure A-2. Behavioral categorization and decision allocation decision tree (continued).

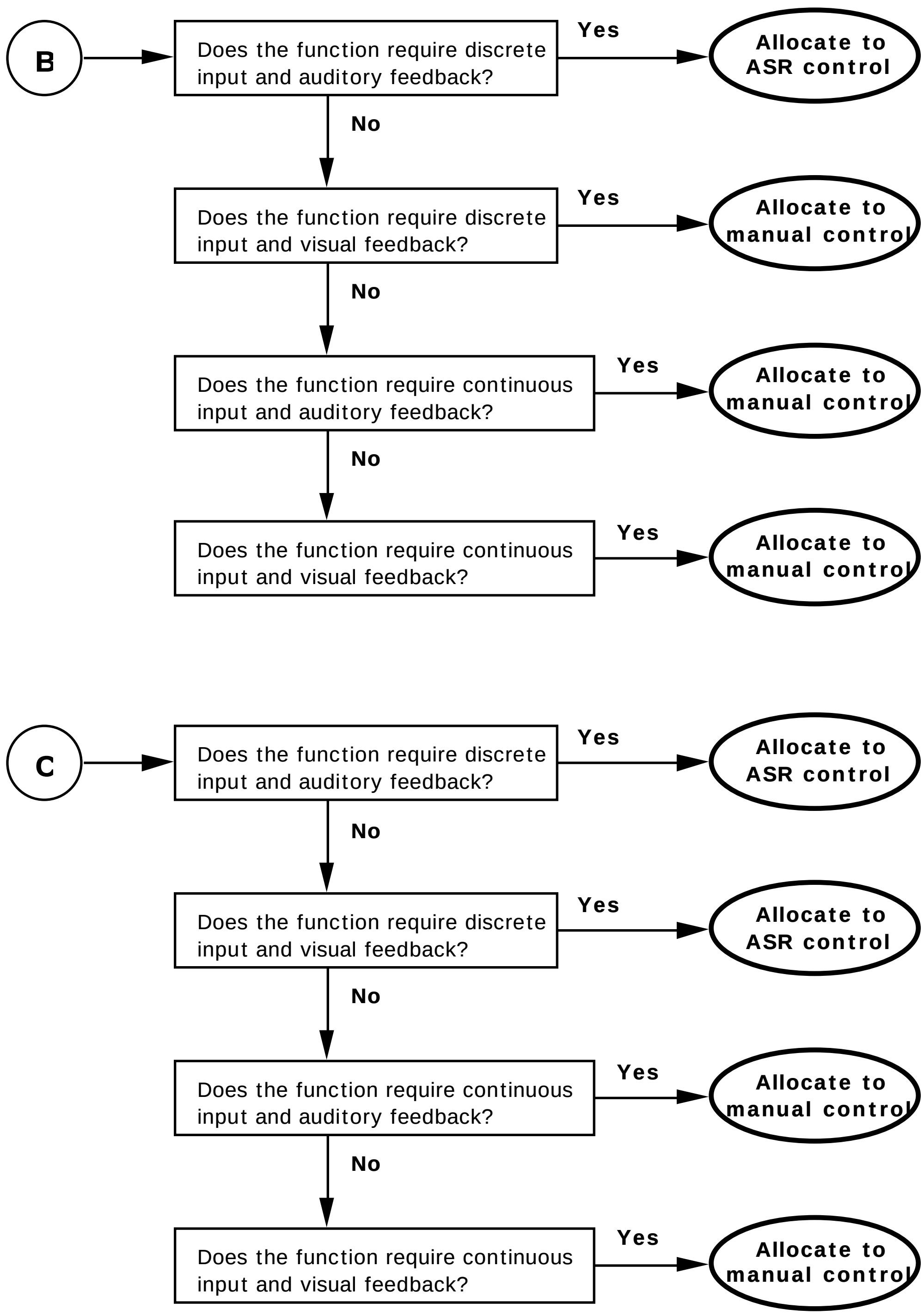

Figure A-2. Behavioral categorization and decision allocation decision tree (continued).

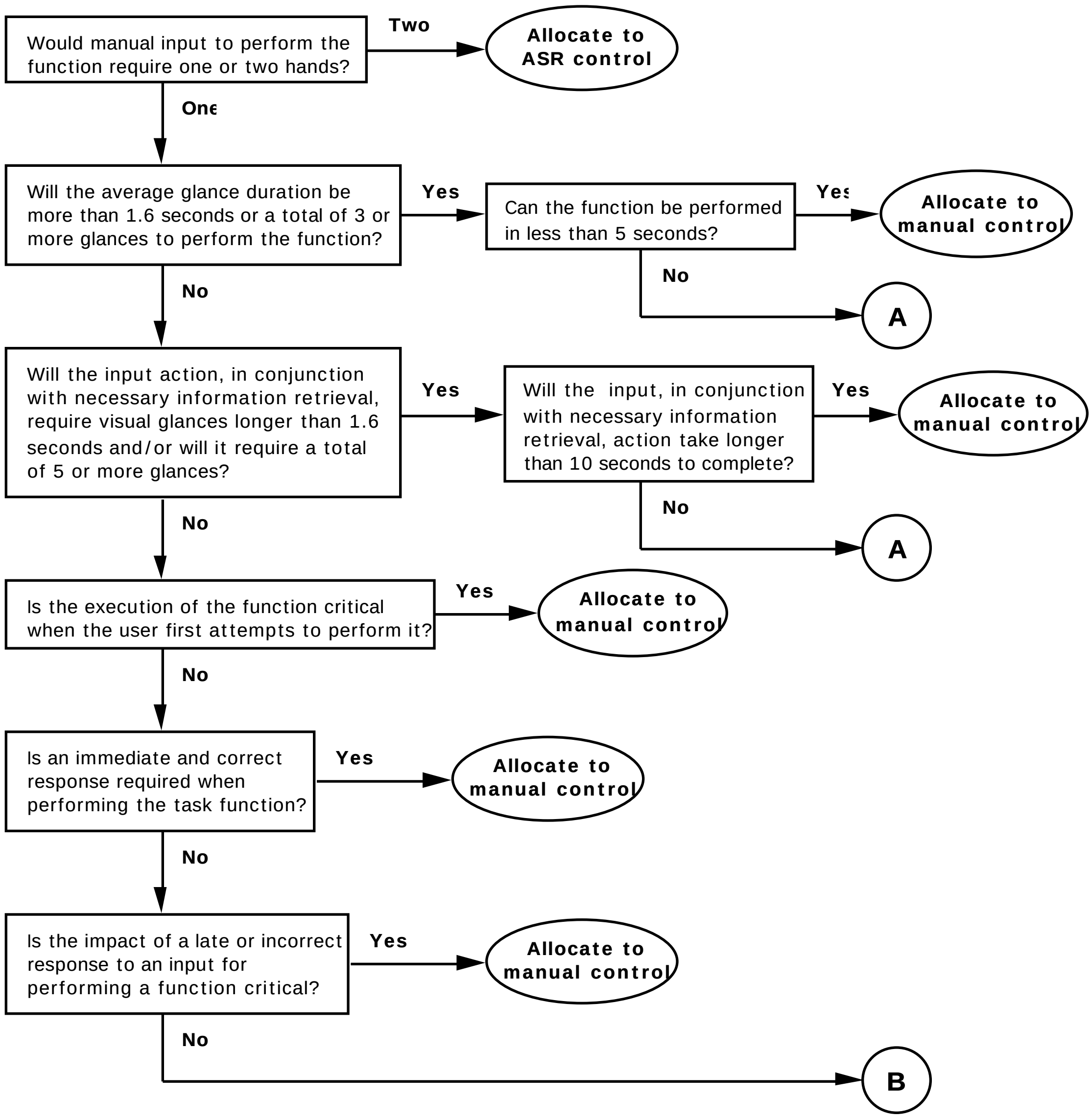

Figure A-3. Safety criteria decision tree.

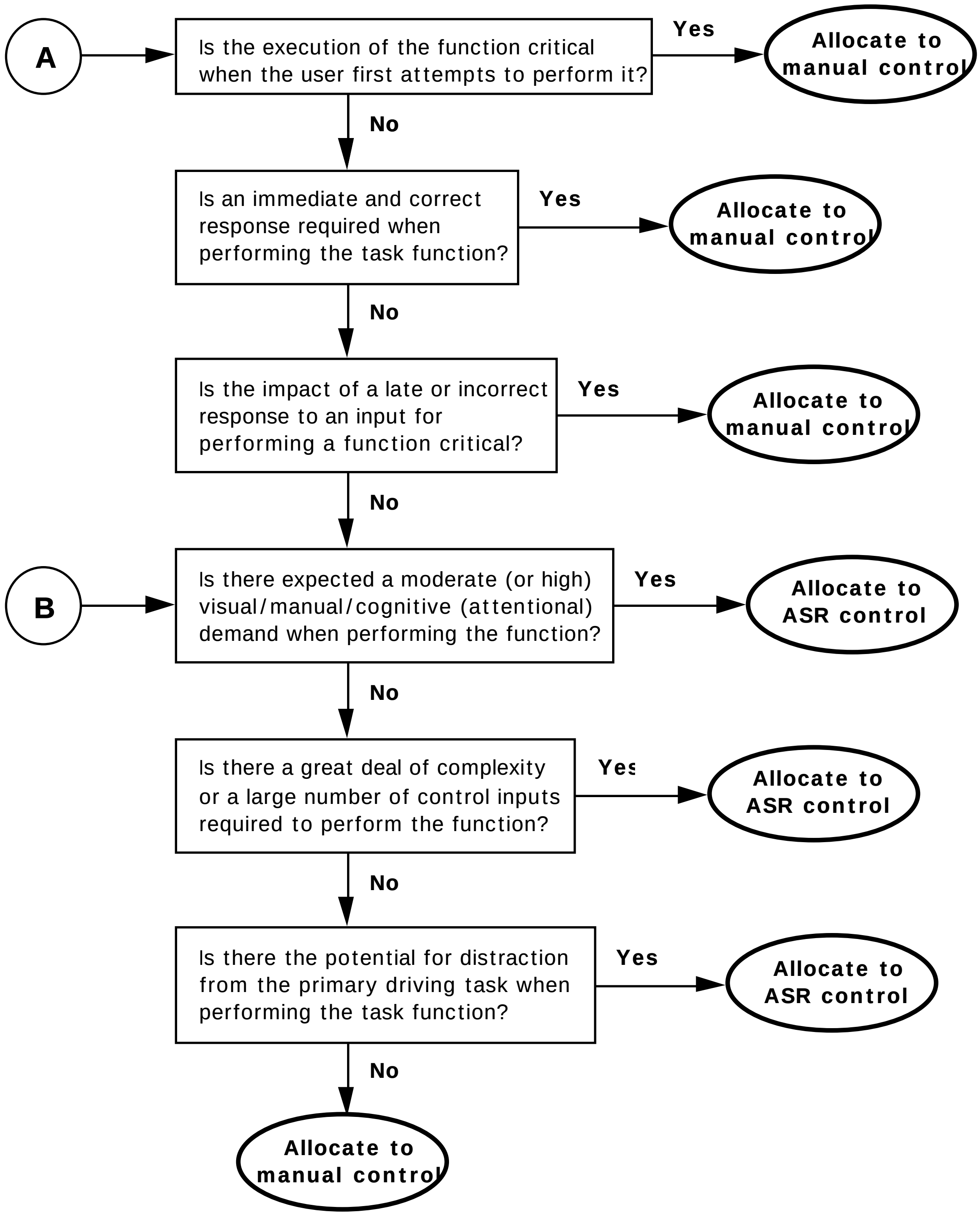

Figure A-3. Safety criteria decision tree (continued).

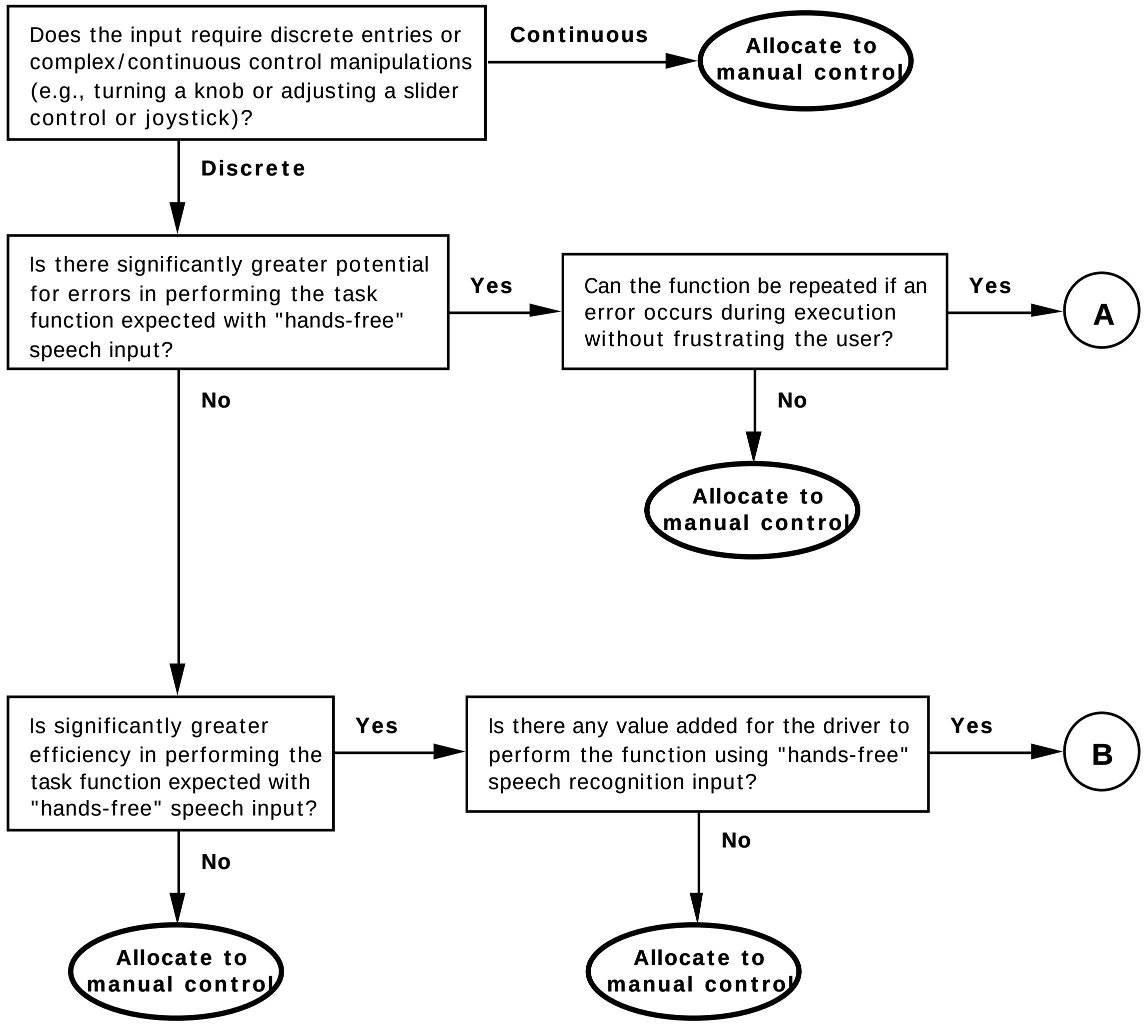

Figure A-4. Usability criteria decision tree.

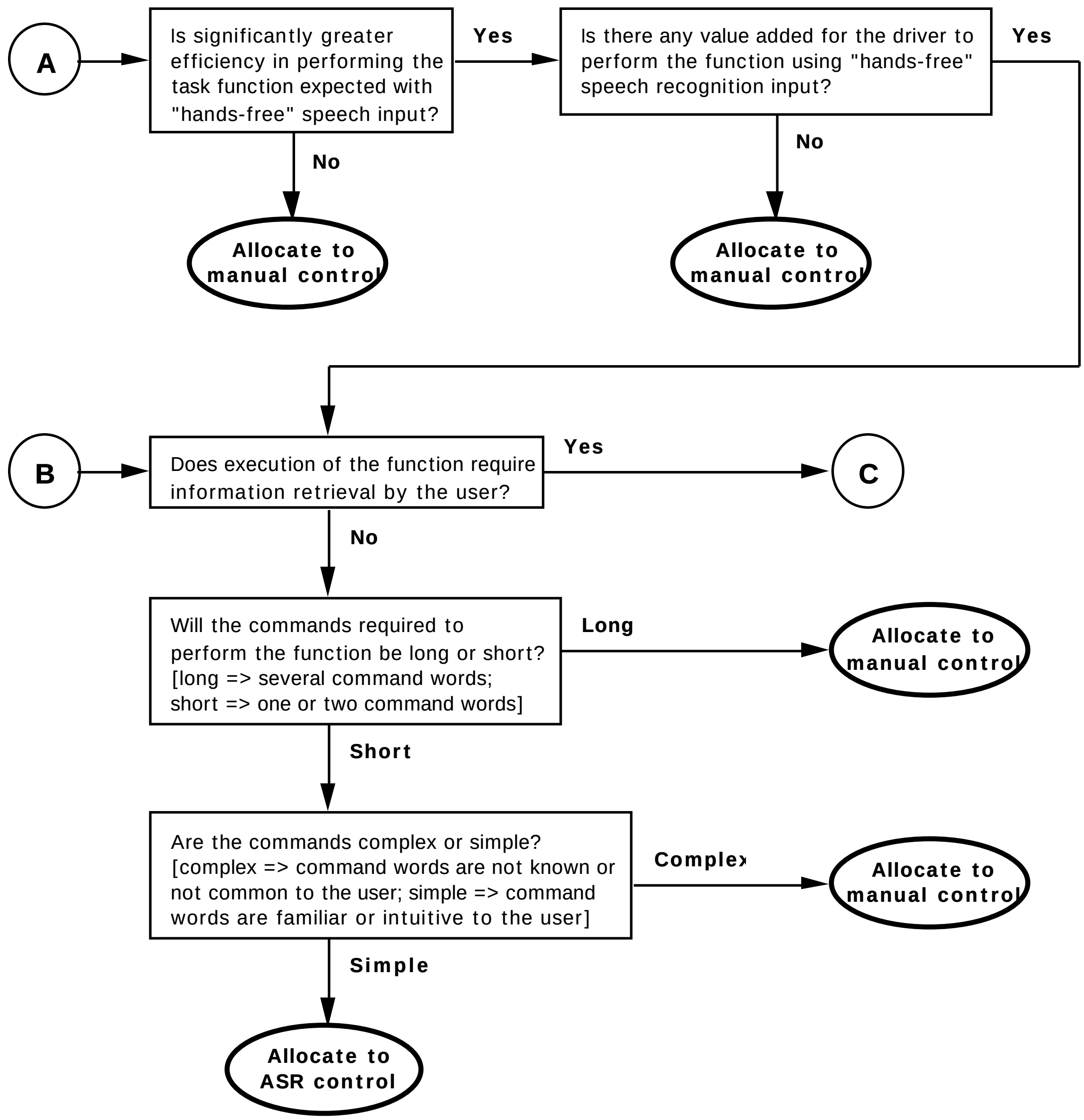

Figure A-4. Usability criteria decision tree (continued).

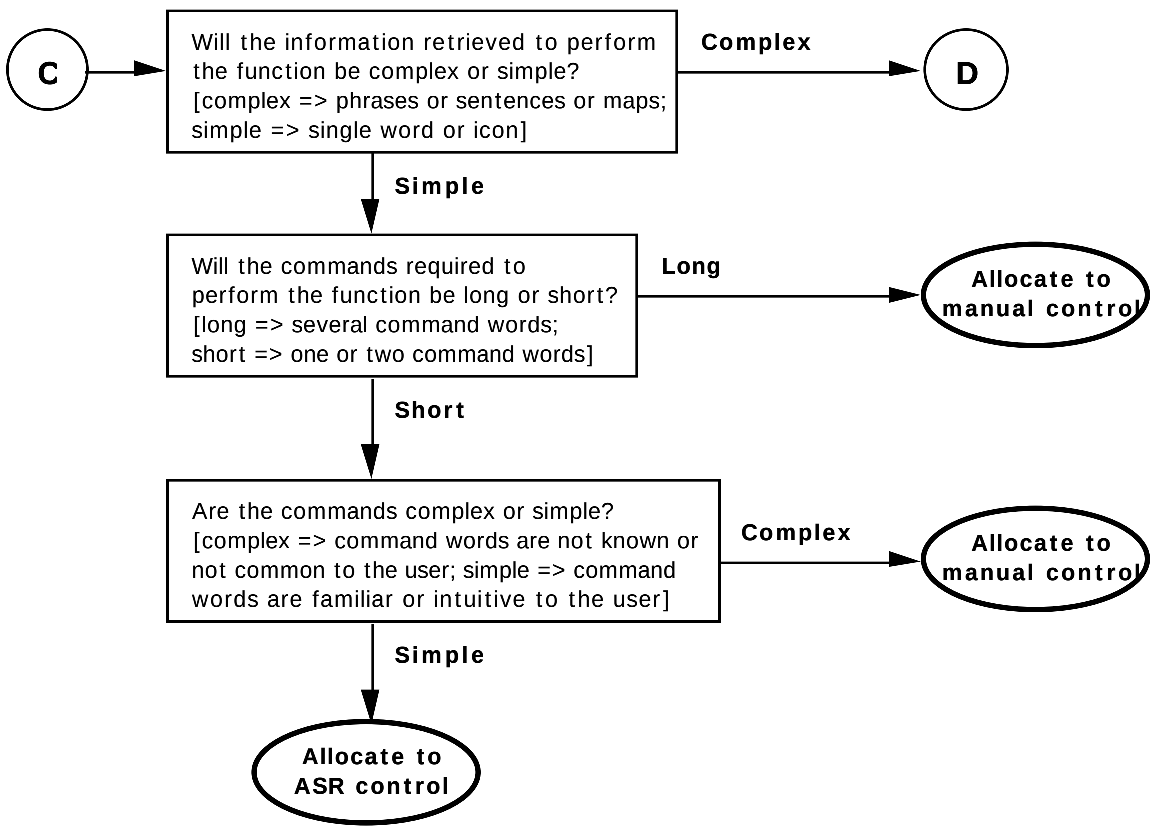

Figure A-4. Usability criteria decision tree (continued).

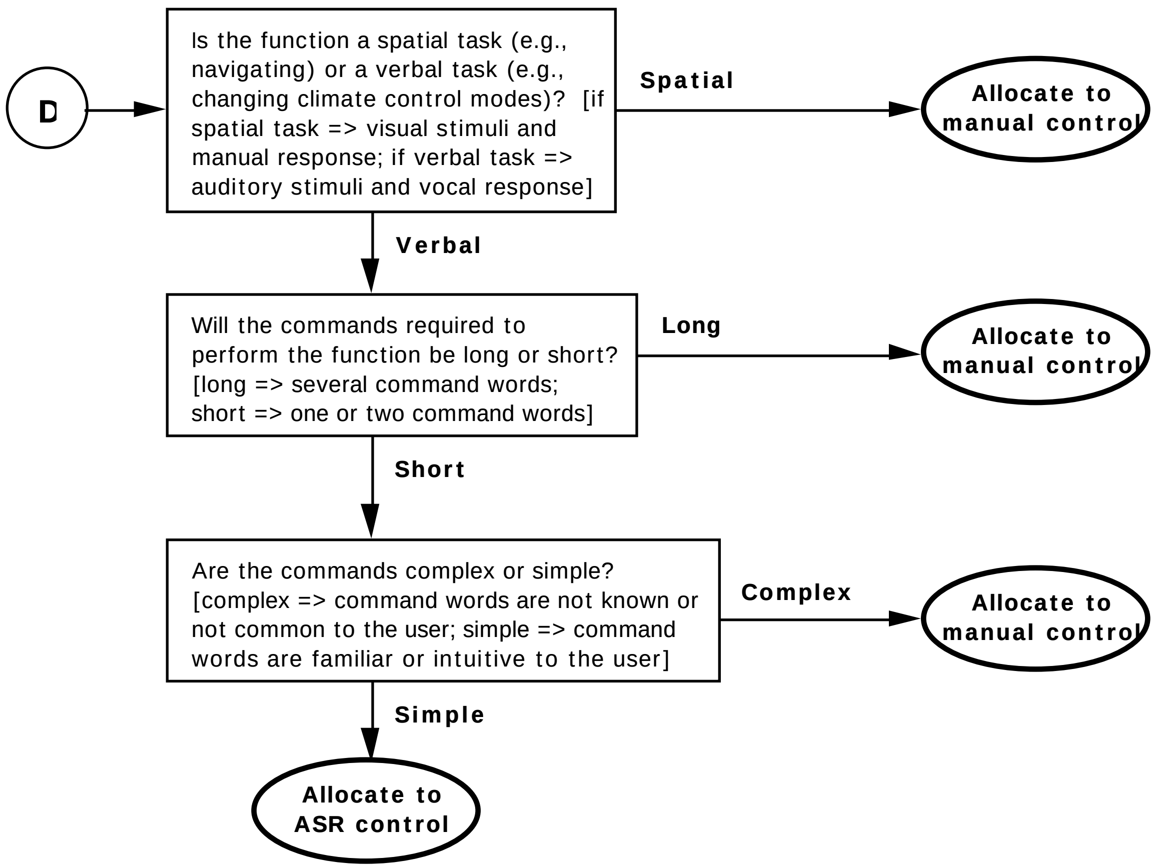

Figure A-4. Usability criteria decision tree (continued).

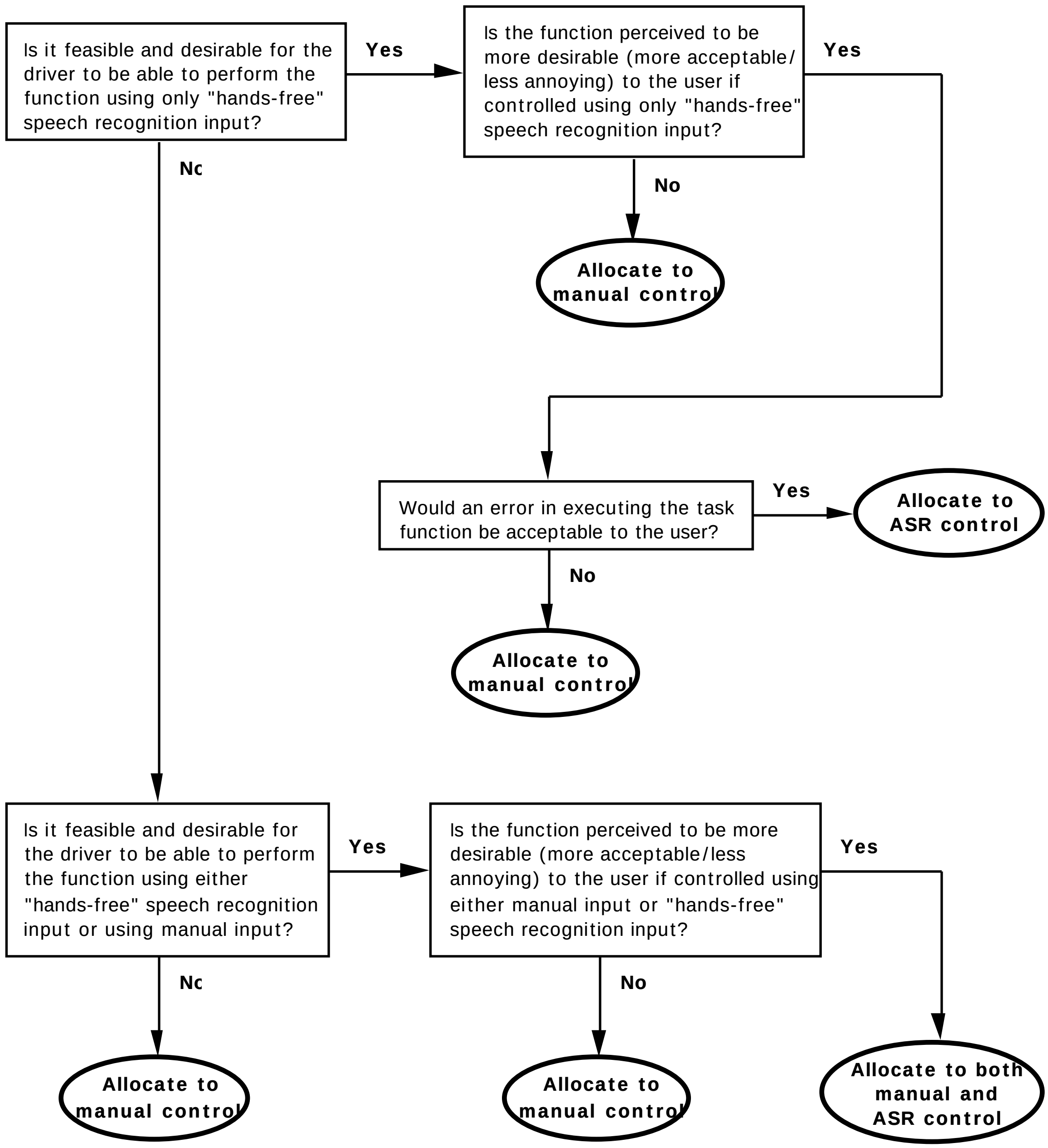

Figure A-5. Driver preference/acceptance criteria decision tree.

APPENDIX B: Task-Function Input-Modality Allocation Decision Chart for Selected In-Vehicle Tasks Performed in Experiments 1 and 2

TASK-FUNCTION ALLOCATION DECISION CHART FOR THE IN-VEHICLE TASKS PERFORMED IN THE AURORA DURING THE TWO EXPERIMENTS

System	Task Function	Driving-Function Category	Behavioral Category	Behavioral Allocation Decision	Safety Allocation Decision	Usability Allocation Decision	Driver Preference/ Acceptance Allocation Decision	**Final Allocation Decision**
Climate	Adjust fan speed	Normal	Simple Locate and Reach/ Gross Adjustment	ASR control	ASR control	Manual control	Both ASR and manual control	**ASR**
Climate	Adjust temperature	Normal	Simple Locate and Reach/Fine Adjustment	Manual control	ASR control	Manual control	Both ASR and manual control	**Manual**
Climate	Select climate mode	Normal	Simple Locate and Reach/ Gross Adjustment	ASR control	ASR control	Manual control	Both ASR and manual control	**ASR**
Audio	Adjust volume	Luxury	Simple Locate and Reach/Fine Adjustment	Manual control	Manual control	Manual control	Manual control	**Manual**
Audio	Tune to a radio station	Luxury	Simple Locate and Reach/Fine Adjustment	Manual control	ASR control	Manual control	Both ASR and manual control	**Manual**
Audio	Select a preset radio station (5)	Luxury	Selective Activation	ASR control	ASR control	ASR control	Both ASR and manual control	**ASR**
Audio	Seek "up" the radio frequencies	Luxury	Simple Locate and Reach/ Gross Adjustment	ASR control	ASR control	Manual control	Both ASR and manual control	**ASR**
DIC	Display average or current fuel economy	Luxury	Selective Activation	ASR control	ASR control	ASR control	ASR control	**ASR**
DIC	Display distance to destination	Luxury	Selective Activation	ASR control	ASR control	ASR control	ASR control	**ASR**

System	Task Function	Driving-Function Category	Behavioral Category	Behavioral Allocation Decision	Safety Allocation Decision	Usability Allocation Decision	Driver Preference/ Acceptance Allocation Decision	**Final Allocation Decision**
DIC	Reset displayed mode	Luxury	Selective Activation	ASR control	ASR control	ASR control	ASR control	**ASR**
DIC	Turn off system	Luxury	Selective Activation	ASR control	ASR control	ASR control	ASR control	**ASR**
Misc.	Lock or unlock doors	Normal	Simple Locate and Reach	ASR control	Manual control	Manual control	Both ASR and manual control	**Manual**
Misc.	Activate/ deactivate cruise control	Normal	Simple Locate and Reach	ASR control	Manual control	Manual control	Both ASR and manual control	**Manual**
Misc.	Set cruise speed (cruise control)	Normal	Simple Locate and Reach	ASR control	Manual control	ASR control	Both ASR and manual control	**ASR**
Misc.	Resume cruise speed (cruise control)	Normal	Simple Locate and Reach	ASR control	Manual control	Manual control	Both ASR and manual control	**Manual**
Misc.	Activate turn signal	Normal	Simple Locate and Reach	ASR control	Manual control	Manual control	Manual control	**Manual**
Misc.	Select windshield wiper operating mode	Normal	Simple Locate and Reach/ Gross Adjustment	Manual control	Manual control	Manual control	Manual control	**Manual**
Cell Phone	Dial phone number	Luxury	Selective Activation/ Gross Adjustment	ASR control	ASR control	ASR control	ASR control	**ASR**
Cell Phone	Dial phone number from preset memory	Luxury	Selective Activation/ Gross Adjustment	ASR control	ASR control	ASR control	ASR control	**ASR**
Cell Phone	End phone call (hang up)	Luxury	Selective Activation	ASR control	ASR control	ASR control	ASR control	**ASR**

APPENDIX C: Informed Consent for Experiment 1

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Informed Consent for Participants of Investigative Projects

Title of Project: In-vehicle Task Performance Study -- Experiment 1
Investigators: Andrew W. Gellatly, Industrial and Systems Engineering graduate student
Dr. Thomas A. Dingus, Industrial and Systems Engineering Professor and Director of the Virginia Tech Center for Transportation Research

I. The Purpose of this Research Project

The purpose of this experiment is to evaluate driving behavior and performance while subjects concurrently perform in-vehicle tasks using a prototype speech recognition control system. These tasks will include adjusting the audio system; adjusting the climate-control system; using the cellular telephone; and accessing the driver-information system. The data obtained will be used to evaluate the effectiveness of speech input for controlling in-vehicle tasks and improve the design of secondary automotive control systems. Ten subjects, each tested individually, will participate in this experiment.

II. Procedures

In the study, you will be asked to perform specific in-car tasks as you drive on U.S. Route 460, west of Blacksburg (that is, toward West Virginia). Two trained experimenters will ride in the research vehicle with you during the experiment to assist in the data-gathering process and to help ensure the safe operation of the experimental vehicle. It is your responsibility as the driver to obey all traffic regulations and to maintain safe operation of the vehicle at all times. You must treat the driving task as the primary task and perform the other instructed tasks only when it is safe to do so. You will be required to have the lap/shoulder belt restraint system securely fastened while driving.

The experimental vehicle is a late model American car. The car is equipped with an automatic transmission, analog instrument cluster, and electronic audio, climate-control, and driver-information systems. The car is also equipped with a speech recognition system. In this study, you will drive and perform in-vehicle tasks using the speech recognition system.

The vehicle is also outfitted with devices designed to monitor various relevant aspects of your driving behavior (for example, video cameras and recorder, microphones, and computers). These measurement devices do not require that your attention be diverted from the driving task. All equipment will be placed in the vehicle and secured such that they will not present a hazard. Also, a fire extinguisher, a first-aid kit, and a cellular phone will be carried in the vehicle at all times, in case of an emergency.

The study will consist of four experimental stages. The experimental stages will proceed as follows.

1. Introductory Stage (only performed on Day 1 prior to the experimental session)

This stage consists of preliminaries. You will thoroughly read the informed consent form. Assuming that you sign the informed consent form, we will ask you to fill out a brief medical screening questionnaire. Next, we will give you a simple vision test and we will also ask to see your driver's license. Once you successfully complete all of these preliminaries, we will begin your training. The first stage is expected to last about 10 minutes.

2. Training Stage (performed each day of the experimental sessions prior to the driving stage)

We will take you to the research vehicle where we will train you on the use of the speech recognition system. Since the instrument panels and controls may differ from the vehicle you currently drive, it is necessary to train you on the in-vehicle tasks you will be performing throughout the experiment. We will then ask you to perform a series of tasks using the speech recognition system you were just trained on. You will also be asked several questions about the tasks which you performed.

This part of the study will be performed both with the car parked in an off-road location and while driving on U.S. Route 460. This stage is expected to last approximately 30 minutes.

3. Driving Stage (performed each day of the experimental sessions)

After a short rest break, you will drive the vehicle to U.S. Route 460. Shortly after reaching the designated speed on Route 460, you will be asked to begin performing a series of instructed in-vehicle tasks. The driving stage will alternate between periods of regular driving and driving while performing the various tasks for which you have been trained. This stage is expected to last approximately 70 minutes in length depending on the number of experimental conditions and amount of re-training required. At the end of the driving stage, you will return to Blacksburg. Upon returning to Blacksburg after experimental sessions 1 and 2, you are free to leave.

4. Debriefing and Payment Stage (only performed on Day 3 at the end of the experimental session)

On returning to Blacksburg, and after having had a chance to refresh, you will be asked to fill out a questionnaire on the control systems you used and read an experiment debriefing statement. You will then be paid and dismissed. This stage should take about 10 minutes.

Your total participation time will be approximately six hours (two hours per experimental session over three sessions/days), but may be somewhat shorter or longer depending on length of rest breaks and amount of training needed.

If during the study you feel that you cannot continue for any reason, you have the right to terminate your participation; you will be paid in full for the amount of time you participated. This includes the right to withdraw at any time after having read and signed the informed consent form. If you withdraw during the driving stage, the experimenter will take over the driving and return you to Blacksburg.

If you have any questions about the experiment or your rights as a participant after reading the informed consent form, please do not hesitate to ask. We will answer your questions as openly and honestly as possible.

III. Risks

There are some risks and discomforts to which subjects are exposed in volunteering for this research. The risks are:

(1) The risk of an accident normally associated with driving an automobile in light or moderate traffic, as well as on straight and curved roadways.
(2) The slight additional risk of an accident that might possibly occur while performing instructed in-vehicle tasks. Past research indicates that this risk is minimal.
(3) Possible fatigue due to the length of the experiment. However, you will be given short rest breaks during the experimental session.

(4) While you are driving the vehicle, you will be videotaped by cameras. Due to this fact, we will ask you not to wear sunglasses. If this at any time during the course of the experiment impairs your ability to drive the vehicle safely, you should notify the experimenter.

The following precautions will be taken to ensure minimal risk to the subjects:

(1) The experimenter will monitor your driving, and will ask you to stop if he feels the risks are too great to continue. However, as long as you are driving the research vehicle, it remains your responsibility to drive in a safe, legal manner.

(2) You will be required to wear the lap and shoulder belt restraint system anytime the car is on the road. The vehicle is also equipped with a driver's side airbag supplemental restraint system.

(3) The vehicle is equipped with a fire extinguisher, first-aid kit, and a cellular phone.

(4) If an accident does occur, the experimenter will arrange medical transportation to a nearby hospital emergency room. You will be required to undergo examination by medical personnel in the emergency room.

IV. Benefits of this Research Project

While there are no direct benefits to you from this research (other than payment), you may find the experiment interesting. No promise or guarantee of benefits have been made to encourage you to participate. Your participation, along with that of the other volunteers, should make it possible to improve the design of automotive secondary-control systems. Improvements in the design of automotive secondary control systems may have a significant impact on driving safety, system usability, and consumer satisfaction.

V. Extent of Anonymity and Confidentiality

The data gathered in this experiment will be treated with confidentiality. Shortly after you have participated, your name will be separated from your data. A coding scheme will be employed to identify your data by gender and subject number only (e.g., Male, Subject No. 3). You have the right to see your data and withdraw it from the study if you so desire. Please inform the experimenter immediately of this decision, as the data will be difficult to track once the session is over.

Eye movement behavior is measured using a video camera and recorder during the experiment. A camera, positioned inside the center rearview mirror, is used to record drivers' eye movements. The video image recorded is of the driver's head with some additional space around the head to accommodate any head-movements by the driver during data collection. The videotapes will be stored in a locked filing cabinet at the Virginia Tech Center for Transportation Research. Access to the tapes will be under the supervision of Dr. Thomas A. Dingus. Andrew W. Gellatly will have access to the tapes and will score the eye movement behavior using "frame-by-frame" analysis. The video tapes will be destroyed three months after the data has been analyzed and the results written up (approximately May of 1997).

At no time will the researchers release the results of the study to anyone other than individuals working on the project without your written consent.

VI. Compensation

You will be paid $5 per hour for the time you actually spend in the experiment. Payment will be made immediately after you have finished your participation.

VII. Freedom to Withdraw

You should know that at any time you are free to withdraw from participation in this research program without penalty. No one will try to make you continue if you do not want to continue, and you will be paid in full for the amount of time you participated.

VIII. Approval of Research

This research project has been approved, as required by the Institutional Review Board for Research Involving Human Subjects at Virginia Polytechnic Institute and State University, by the Department of Industrial and Systems Engineering and the Virginia Tech Center for Transportation Research.

IX. Subject's Responsibilities

I voluntarily agree to participate in this study. I have the following responsibilities:

(1) I should not volunteer for participation in this research if I am younger than 18 years of age or older than 85 years of age, or if I do not have a valid driver's license, or if I am not in good health, or if I am pregnant.

(2) I should not take part in the driving task if I have taken any drug, alcoholic beverage, or medication within the previous 24 hours which might affect my ability to safely operate an automobile. It is my responsibility to inform the experimenters of any additional conditions which might interfere with my ability to drive. Such conditions would include inadequate sleep, hunger, hangover, headache, cold symptoms, depression, allergies, emotional upset, visual or hearing impairment, seizures (fits), nerve or muscle disease, or other similar conditions.

(3) As the driver of the research vehicle, I must obey all traffic regulations and maintain safe operation of the vehicle at all times. I will treat the driving task as the primary task and perform the other instructed tasks only when it is safe to do so.

X. Subject's Permission

I have read and understand the Informed Consent and conditions of this research project. I have had all my questions answered. I hereby acknowledge the above and give my voluntary consent for participation in this project.

If I participate, I may withdraw at any time without penalty. I agree to abide by the rules of this research project.

______________________________ ______________________________

Signature Date

Should I have any questions about this research or its conduct, I may contact:

Andrew W. Gellatly Principal Investigator	231-7740
Dr. Thomas A. Dingus Faculty Advisor	231-8831
E.R. Stout Chair, IRB Research Division	231-9359

APPENDIX D: Informed Consent for Experiment 2

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Informed Consent for Participants of Investigative Projects

Title of Project: In-vehicle Task Performance Study -- Experiment 2
Investigators: Andrew W. Gellatly, Industrial and Systems Engineering graduate student
Dr. Thomas A. Dingus, Industrial and Systems Engineering Professor and Director of the Virginia Tech Center for Transportation Research

I. The Purpose of this Research Project

The purpose of this experiment is to evaluate driving behavior and performance while subjects concurrently perform in-vehicle tasks using conventional types of in-vehicle controls compared with tasks performed using an alternative speech recognition control system. These tasks will include adjusting the audio system; adjusting the climate-control system; using the cellular telephone; and accessing the driver-information system. The data obtained will be used to evaluate the effectiveness of speech recognition controls, as compared with conventional manual controls for performing in-vehicle tasks. Improvements in driving safety and performance while performing in-vehicle tasks may be realized from the results. Twelve subjects, each tested individually, will participate in this experiment.

II. Procedures

In the study, you will be asked to perform specific in-car tasks as you drive on U.S. Route 460, west of Blacksburg (that is, toward West Virginia). Two trained experimenter will ride in the research vehicle with you during the experiment to assist in the data-gathering process and to help ensure the safe operation of the experimental vehicle. It is your responsibility as the driver to obey all traffic regulations and to maintain safe operation of the vehicle at all times. You must treat the driving task as the primary task and perform the other instructed tasks only when it is safe to do so. You will be required to have the lap/shoulder belt restraint system securely fastened while driving.

The experimental vehicle is a late model American car. The car is equipped with an automatic transmission, analog instrument cluster, and electronic audio, climate-control and driver-information systems. The car is also equipped with a speech recognition system. In some parts of the study, you will drive and perform in-vehicle tasks using conventional control inputs, and in other parts of the study you will drive and perform in-vehicle tasks using the speech recognition system.

The vehicle is also outfitted with devices designed to monitor various relevant aspects of your driving behavior (for example, video cameras and recorder, microphones, and computers). These measurement devices do not require that your attention be diverted from the driving task. All equipment will be placed in the vehicle and secured such that they will not present a hazard. Also, a fire extinguisher, a first-aid kit, and a cellular phone will be carried in the vehicle at all times, in case of an emergency.

The study will consist of four experimental stages. The experimental stages will proceed as follows over the three days of data collection.

1. Introductory Stage (only performed on Day 1 prior to the experimental session)

This stage consists of preliminaries. You will thoroughly read the informed consent form. Assuming that you sign the informed consent form, we will ask you to fill out a brief medical screening questionnaire. Next, we will give you a simple vision test and we will also ask to see your driver's license. Once you successfully complete all of these preliminaries, we will begin your training. The first stage is expected to last about 15 minutes.

2. Training Stage (performed each day of the experimental sessions prior to the driving stage)

Depending on the group to which you have been assigned, you will experience either the manual controlled in-vehicle tasks or one of the alternative speech recognition controlled in-vehicle tasks. We will take you to the research vehicle where we will train you on the use of either the manual controlled tasks or the speech controlled tasks. Since the instrument panels and controls may differ from the vehicle you currently drive, it is necessary to train you on the in-vehicle tasks you will be performing. We will then ask you to perform a series of tasks using the appropriate control system you were just trained on. You will also be asked several questions about the tasks which you performed.

This part of the study will be performed both with the car parked in an off-road location and while driving on U.S. Route 460. This stage is expected to last approximately 30 minutes.

3. Driving Stage (performed each day of the experimental sessions)

After a short rest break, you will drive the vehicle to U.S. Route 460. Shortly after reaching the designated speed on Route 460, you will be asked to begin performing a series of instructed tasks. The driving session will alternate between periods of regular driving and driving while performing the various tasks for which you have been trained. This stage is expected to last approximately 75 minutes. At the end of the stage, you will return to Blacksburg. Upon returning to Blacksburg after experimental sessions 1 and 2, you are free to leave.

4. Debriefing and Payment Stage (only performed on Day 3 at the end of the experimental session)

On returning to Blacksburg, and after having had a chance to refresh, you will be asked to fill out a questionnaire on the control systems you used and read an experiment debriefing statement. You will then be paid and dismissed. This stage should take about 10 minutes.

Your total participation time will be approximately 7 hours (2.3 hours per experimental session over 3 sessions/days), but may be somewhat shorter or longer depending on length of rest breaks and amount of training needed during the experimental sessions.

If during the study you feel that you cannot continue for any reason, you have the right to terminate your participation; you will be paid in full for the amount of time you participated. This includes the right to withdraw at any time after having read and signed the informed consent form. If you withdraw during the driving stage, the experimenter will take over the driving and return you to Blacksburg.

If you have any questions about the experiment or your rights as a participant after reading the informed consent form, please do not hesitate to ask. We will answer your questions as openly and honestly as possible.

III. Risks

There are some risks and discomforts to which subjects are exposed in volunteering for this research. The risks are:

(1) The risk of an accident normally associated with driving an automobile in light or moderate traffic, as well as on straight and curved roadways.

(2) The slight additional risk of an accident that might possibly occur while performing instructed in-vehicle tasks. Past research indicates that this risk is minimal.
(3) Possible fatigue due to the length of the experiment. However, you will be given short rest breaks during the experimental session.
(4) While you are driving the vehicle, you will be videotaped by cameras. Due to this fact, we will ask you not to wear sunglasses. If this at any time during the course of the experiment impairs your ability to drive the vehicle safely, you should notify the experimenter.

The following precautions will be taken to ensure minimal risk to the subjects:

(1) The experimenter will monitor your driving, and will ask you to stop if he feels the risks are too great to continue. However, as long as you are driving the research vehicle, it remains your responsibility to drive in a safe, legal manner.
(2) You will be required to wear the lap and shoulder belt restraint system anytime the car is on the road. The vehicle is also equipped with a driver's side airbag supplemental restraint system.
(3) The vehicle is equipped with a fire extinguisher, first-aid kit, and a cellular phone.
(4) If an accident does occur, the experimenter will arrange medical transportation to a nearby hospital emergency room. You will be required to undergo examination by medical personnel in the emergency room.

IV. Benefits of this Research Project

While there are no direct benefits to you from this research (other than payment), you may find the experiment interesting. No promise or guarantee of benefits have been made to encourage you to participate. Your participation, along with that of the other volunteers, should make it possible to improve the design of automotive secondary-control systems. Improvements in the design of automotive secondary control systems may have a significant impact on driving safety, system usability, and consumer satisfaction.

V. Extent of Anonymity and Confidentiality

The data gathered in this experiment will be treated with confidentiality. Shortly after you have participated, your name will be separated from your data. A coding scheme will be employed to identify your data by gender and subject number only (e.g., Male, Subject No. 3). You have the right to see your data and withdraw it from the study if you so desire. Please inform the experimenter immediately of this decision, as the data will be difficult to track once the session is over.

Eye movement behavior is measured using a video camera and recorder during the experiment. A camera, positioned inside the center rearview mirror, is used to record drivers' eye movements. The video image recorded is of the driver's head with some additional space around the head to accommodate any head-movements by the driver during data collection. The videotapes will be stored in a locked filing cabinet at the Virginia Tech Center for Transportation Research. Access to the tapes will be under the supervision of Dr. Thomas A. Dingus. Andrew W. Gellatly will have access to the tapes and will score the eye movement behavior using "frame-by-frame" analysis. The video tapes will be destroyed three months after the data has been analyzed and the results written up (approximately May of 1997).

At no time will the researchers release the results of the study to anyone other than individuals working on the project without your written consent.

VI. Compensation

You will be paid $10 per hour ($15 per hour for older drivers) for the time you actually spend in the experiment. Payment will be made immediately after you have finished your participation.

VII. Freedom to Withdraw

You should know that at any time you are free to withdraw from participation in this research program without penalty. No one will try to make you continue if you do not want to continue, and you will be paid in full for the amount of time you participated.

VIII. Approval of Research

This research project has been approved, as required by the Institutional Review Board for Research Involving Human Subjects at Virginia Polytechnic Institute and State University, by the Department of Industrial and Systems Engineering and the Virginia Tech Center for Transportation Research.

IX. Subject's Responsibilities

I voluntarily agree to participate in this study. I have the following responsibilities:

(1) I should not volunteer for participation in this research if I am younger than 18 years of age or older than 85 years of age, or if I do not have a valid driver's license, or if I am not in good health, or if I am pregnant.

(2) I should not take part in the driving task if I have taken any drug, alcoholic beverage, or medication within the previous 24 hours which might affect my ability to safely operate an automobile. It is my responsibility to inform the experimenters of any additional conditions which might interfere with my ability to drive. Such conditions would include inadequate sleep, hunger, hangover, headache, cold symptoms, depression, allergies, emotional upset, visual or hearing impairment, seizures (fits), nerve or muscle disease, or other similar conditions.

(3) As the driver of the research vehicle, I must obey all traffic regulations and maintain safe operation of the vehicle at all times. I will treat the driving task as the primary task and perform the other instructed tasks only when it is safe to do so.

X. Subject's Permission

I have read and understand the Informed Consent and conditions of this research project. I have had all my questions answered. I hereby acknowledge the above and give my voluntary consent for participation in this project.

If I participate, I may withdraw at any time without penalty. I agree to abide by the rules of this research project.

____________________________ ____________________________

Signature Date

Should I have any questions about this research or its conduct, I may contact:

Andrew W. Gellatly 231-7740
Principal Investigator

Dr. Thomas A. Dingus 231-8831
Faculty Advisor

E.R. Stout 231-9359
Chair, IRB Research Division

APPENDIX E: Health Screening Questionnaire

IN-VEHICLE TASK PERFORMANCE STUDY

Health Screening Questionnaire

1. Are you in good general health? Yes No

   If no, list any health-related conditions you are experiencing or have experienced in the recent past.

   ____

   ____

   ____

2. Have you, in the last 24 hours, experienced any of the following conditions?

Inadequate sleep	Yes	No
Unusual hunger	Yes	No
Hangover	Yes	No
Headache	Yes	No
Cold symptoms	Yes	No
Depression	Yes	No
Allergies	Yes	No
Emotional upset	Yes	No

3. Do you have a history of any of the following?

   Visual Impairment Yes No

   (If yes, please describe.)

   ____

   ____

   ____

   Hearing Impairment Yes No

   (If yes, please describe.)

   ____

   ____

   ____

   Seizures or other lapses of consciousness Yes No

   (If yes, please describe.)

   ____

   ____

   ____

Any disorders similar to the
above or that would impair
your driving ability Yes No

(If yes, please describe.)

____
____
____

4. List any prescription or non-prescription drugs you are currently taking or have taken in the last 24 hours.

____
____

5. List the approximate amount of alcohol (beer, wine, fortified wine, or liquor) you have consumed in the last 24 hours.

____
____

6. Are you taking any drugs of any kind other than those listed in 4 or 5 above?

Yes No

7. If you are female, are you pregnant? Yes No

____________________ ____________________
Signature Date

APPENDIX F: Driver Preference/Acceptance Questionnaire for Experiment 1

IN-VEHICLE TASK PERFORMANCE STUDY

Driver Preference/Acceptance Questionnaire

After subjects perform each in-vehicle task, the experimenter will ask subjects to provide some feedback about the task they just performed. This feedback will be in the form of responses to three five-point rating scales.

The first scale used will be this one:

1	**2**	**3**	**4**	**5**
Very Difficult	Difficult	Neither Easy nor Difficult	Easy	Very Easy

For each in-vehicle task, the experimenter will ask subjects to give a number between 1 and 5 indicating how <u>easy</u> it is to perform the task while driving the vehicle. 1 means that subjects feel it is very difficult to perform the task while driving, and 5 means that subjects feel it is very easy to perform the task while driving. 3 means that subjects feel it is neither difficult nor easy to perform the task while driving.

For example, if the experimenter asks subjects how easy it is to tune the radio to a certain station while driving, subjects will give a lower number if they feel it is not particularly easy, and a higher number if they feel it is particularly easy. Subjects may use any number between 1 and 5.

The second scale used will be this one:

1	**2**	**3**	**4**	**5**
Very Uncomfortable	Uncomfortable	Neither Comfortable nor Uncomfortable	Comfortable	Very Comfortable

For each in-vehicle task, the experimenter will ask subjects to give a number between 1 and 5 indicating how <u>comfortable</u> they feel performing the task while driving the vehicle. 1 means that subjects feel very uncomfortable performing the task while driving, and 5 means that subjects feel very comfortable performing the task while driving. 3 means that subjects feel neither uncomfortable nor comfortable performing the task while driving.

For example, if the experimenter asks subjects how comfortable they feel tuning the radio to a certain station while driving, subjects will give a lower number if they feel very uncomfortable, and a higher number if they feel very comfortable. Subjects may use any number between 1 and 5.

The last scale used will be this one:

1	**2**	**3**	**4**	**5**
Very Distracting	Distracting	Neither Distracting nor Not Distracting	Not Distracting	Not at All Distracting

For each in-vehicle task, the experimenter will ask subjects to give a number between 1 and 5 indicating how <u>distracting</u> subjects feel it is to perform the task while driving the vehicle. In other words, the experimenter will be asking subjects to indicate how much they think the task distracts them from the primary task of driving the vehicle. 1 means that subjects feel that is very distracting to perform the task while driving, and 5 means that subjects feel that it is not distracting at all to perform the task while driving. 3 means that subjects feel it is neither distracting nor not distracting to perform the task while driving.

For example, if the experimenter asks subjects how distracting they feel tuning the radio to a certain station is while driving, subjects will give a lower number if they feel it was very distracting, and a higher number if they feel it was not distracting at all. Subjects may use any number between 1 and 5.

APPENDIX G: Driver Preference/Acceptance Questionnaire for Experiment 2

IN-VEHICLE TASK PERFORMANCE STUDY

Driver Preference/Acceptance Questionnaire 1

After subjects perform each in-vehicle task, the experimenter will ask subjects to provide some feedback about the task they just performed. This feedback will be in the form of responses to three seven-point rating scales.

The first scale used will be this one:

1	**2**	**3**	**4**	**5**	**6**	**7**
Very Difficult			Neither Easy nor Difficult			Very Easy

For each in-vehicle task, the experimenter will ask subjects to give a number between 1 and 7 indicating how <u>easy</u> it is to perform the task while driving the vehicle. 1 means that subjects feel it is very difficult to perform the task while driving, and 7 means that subjects feel it is very easy to perform the task while driving. 4 means that subjects feel it is neither difficult nor easy to perform the task while driving.

For example, if the experimenter asks subjects how easy it is to tune the radio to a certain station while driving, subjects will give a lower number if they feel it is not particularly easy, and a higher number if they feel it is particularly easy. Subjects may use any number between 1 and 7.

The second scale used will be this one:

1	**2**	**3**	**4**	**5**	**6**	**7**
Very Uncomfortable			Neither Comfortable nor Uncomfortable			Very Comfortable

For each in-vehicle task, the experimenter will ask subjects to give a number between 1 and 7 indicating how <u>comfortable</u> they feel performing the task while driving the vehicle. 1 means that subjects feel very uncomfortable performing the task while driving, and 7 means that subjects feel very comfortable performing the task while driving. 4 means that subjects feel neither comfortable nor uncomfortable performing the task while driving.

For example, if the experimenter asks subjects how comfortable they feel tuning the radio to a certain station while driving, subjects will give a lower number if they feel very uncomfortable, and a higher number if they feel very comfortable. Subjects may use any number between 1 and 7.

The last scale used will be this one:

1	2	3	4	5	6	7
Very Distracting			Neither Distracting nor Not Distracting			Not at All Distracting

For each in-vehicle task, the experimenter will ask subjects to give a number between 1 and 7 indicating how distracting subjects feel it is to perform the task while driving the vehicle. In other words, the experimenter will be asking subjects to indicate how much they think the task distracts them from the primary task of driving the vehicle. 1 means that subjects feel that is very distracting to perform the task while driving, and 7 means that subjects feel that it is not distracting at all to perform the task while driving. 4 means that subjects feel it is neither distracting nor not distracting to perform the task while driving.

For example, if the experimenter asks subjects how distracting they feel tuning the radio to a certain station is while driving, subjects will give a lower number if they feel it was very distracting, and a higher number if they feel it was not distracting at all. Subjects may use any number between 1 and 7.

APPENDIX H: Analysis of Variance Tables for Experiment 1 Results

Table H-1. Analysis of variance table for driving performance dependent variable, “Steering Variability/Variance,” and independent variables ASR system feedback modality, error correction method, and recognition error type.

Independent Variable	df	MS	F	p
Feedback	2	0.000041	0.60	0.560
Feedback*Subject	18	0.000068		
Error Correction	1	0.000030	0.84	0.384
Error Correction*Subject	9	0.000035		
Error Type	2	0.000484	10.61	0.001
Error Type*Subject	18	0.000046		
Feedback*Error Correction	2	0.000178	1.91	0.179
Feedback*Error Correction*Subject	18	0.000094		
Feedback*Error Type	4	0.000048	1.25	0..09
Feedback*Error Type*Subject	36	0.000039		
Error Correction*Error Type	2	0.000056	0.71	0.506
Error Correction*Error Type*Subject	18	0.000080		
Feedback*Error Correction*Error Type	4	0.000096	1.54	0.215
Feedback*Error Correction*Error Type*Subject	35	0.000062		

Table H-2. Analysis of variance table for driving performance dependent variable, "Steering Reversals > 6°," and independent variables ASR system feedback modality, error correction method, and recognition error type.

Independent Variable	df	MS	F	p
Feedback	2	265.901	1.21	0.322
Feedback*Subject	18	220.238		
Error Correction	1	251.539	0.77	0.403
Error Correction*Subject	9	326.786		
Error Type	2	655.935	1.82	0.190
Error Type*Subject	18	359.660		
Feedback*Error Correction	2	371.866	1.97	0.170
Feedback*Error Correction*Subject	18	188.553		
Feedback*Error Type	4	257.511	0.97	0.433
Feedback*Error Type*Subject	36	264.193		
Error Correction*Error Type	2	640.440	1.94	0.173
Error Correction*Error Type*Subject	18	330.698		
Feedback*Error Correction*Error Type	4	178.000	0.66	0.626
Feedback*Error Correction*Error Type*Subject	35	271.047		

Table H-3. Analysis of variance table for driving performance dependent variable, "Steering Reversals > 12°," and independent variables ASR system feedback modality, error correction method, and recognition error type.

Independent Variable	df	MS	F	p
Feedback	2	14.839	1.50	0.251
Feedback*Subject	18	9.923		
Error Correction	1	0.085	0.01	0.914
Error Correction*Subject	9	6.951		
Error Type	2	13.349	1.74	0.204
Error Type*Subject	18	7.664		
Feedback*Error Correction	2	13.996	1.17	0.333
Feedback*Error Correction*Subject	18	11.912		
Feedback*Error Type	4	9.873	0.66	0.627
Feedback*Error Type*Subject	36	15.066		
Error Correction*Error Type	2	3.879	0.42	0.663
Error Correction*Error Type*Subject	18	9.220		
Feedback*Error Correction*Error Type	4	3.280	0.38	0.821
Feedback*Error Correction*Error Type*Subject	34	8.622		

Table H-4. Analysis of variance table for driving performance dependent variable, "Accelerator Position Variability," and independent variables ASR system feedback modality, error correction method, and recognition error type.

Independent Variable	df	MS	F	p
Feedback	2	0.00088	1.73	0.205
Feedback*Subject	18	0.00051		
Error Correction	1	0.000001	0.003	0.9535
Error Correction*Subject	9	0.000283		
Error Type	2	0.00285	9.77	0.001
Error Type*Subject	18	0.00029		
Feedback*Error Correction	2	0.00033	1.21	0.322
Feedback*Error Correction*Subject	18	0.00027		
Feedback*Error Type	4	0.00261	0.94	0.451
Feedback*Error Type*Subject	36	0.00028		
Error Correction*Error Type	2	0.00016	0.83	0.452
Error Correction*Error Type*Subject	18	0.00020		
Feedback*Error Correction*Error Type	4	0.00034	1.87	0.139
Feedback*Error Correction*Error Type*Subject	35	0.00018		

Table H-5. Analysis of variance table for driving performance dependent variable, “Mean Velocity,” and independent variables ASR system feedback modality, error correction method, and recognition error type.

Independent Variable	df	MS	F	p
Feedback	2	317.043	2.41	0.118
Feedback*Subject	18	131.720		
Error Correction	1	2.278	0.07	0.792
Error Correction*Subject	9	30.970		
Error Type	2	50.104	1.76	0.201
Error Type*Subject	18	28.546		
Feedback*Error Correction	2	207.027	1.49	0.254
Feedback*Error Correction*Subject	18	139.106		
Feedback*Error Type	4	8.708	0.46	0.767
Feedback*Error Type*Subject	36	19.076		
Error Correction*Error Type	2	14.037	0.38	0.686
Error Correction*Error Type*Subject	18	36.488		
Feedback*Error Correction*Error Type	4	13.675	0.43	0.783
Feedback*Error Correction*Error Type*Subject	35	31.477		

Table H-6. Analysis of variance table for driving performance dependent variable, "Velocity Variability/Variance," and independent variables ASR system feedback modality, error correction method, and recognition error type.

Independent Variable	df	MS	F	p
Feedback	2	184.919	0.81	0.462
Feedback*Subject	18	229.674		
Error Correction	1	78.216	0.43	0.530
Error Correction*Subject	9	183.208		
Error Type	2	1909.996	19.77	0.0001
Error Type*Subject	18	96.605		
Feedback*Error Correction	2	635.087	2.17	0.145
Feedback*Error Correction*Subject	18	292.696		
Feedback*Error Type	4	47.554	0.85	0.506
Feedback*Error Type*Subject	36	56.220		
Error Correction*Error Type	2	23.671	0.74	0.491
Error Correction*Error Type*Subject	18	31.970		
Feedback*Error Correction*Error Type	4	173.000	1.57	0.206
Feedback*Error Correction*Error Type*Subject	35	110.475		

Table H-7. Analysis of variance table for driving performance dependent variable, "Brake Activations," and independent variables ASR system feedback modality, error correction method, and recognition error type.

Independent Variable	df	MS	F	p
Feedback	2	3.938	0.15	0.859
Feedback*Subject	18	25.674		
Error Correction	1	17.753	1.48	0.255
Error Correction*Subject	9	12.013		
Error Type	2	103.919	9.54	0.002
Error Type*Subject	18	10.893		
Feedback*Error Correction	2	0.315	0.04	0.9614
Feedback*Error Correction*Subject	18	7.993		
Feedback*Error Type	4	17.664	2.17	0.092
Feedback*Error Type*Subject	36	8.141		
Error Correction*Error Type	2	2.296	0.22	0.804
Error Correction*Error Type*Subject	18	10.424		
Feedback*Error Correction*Error Type	4	9.075	1.29	0.294
Feedback*Error Correction*Error Type*Subject	34	7.031		

Table H-8. Analysis of variance table for driving performance dependent variable, "Reaction Time to Forward-Scene-Event/Light," and independent variables gender and during task/during driving.

Independent Variable	df	MS	F	p
Gender	1	16.294	1.27	0.292
Subject(Gender)	8	12.786		
During Task/During Driving	1	11.681	2.61	0.145
During Task/During Driving*Subject(Gender)	8	4.476		
Gender*During Task/During Driving	1	0.839	0.19	0.676
During Task/During Driving*Subject(Gender)	8	4.476		

Table H-9. Analysis of variance table for driving performance dependent variable, "Reaction Time to Forward-Scene-Event/Light," and independent variables ASR system feedback modality, error correction method, and recognition error type.

Independent Variable	df	MS	F	p
Feedback	2	0.912	0.27	0.769
Feedback*Subject	18	3.416		
Error Correction	1	6.372	2.29	0.165
Error Correction*Subject	9	2.785		
Error Type	2	0.549	0.38	0.690
Error Type*Subject	18	1.447		
Feedback*Error Correction	2	0.250	0.18	0.839
Feedback*Error Correction*Subject	17	1.403		
Feedback*Error Type	4	1.956	1.35	0.270
Feedback*Error Type*Subject	36	1.446		
Error Correction*Error Type	2	0.343	0.16	0.853
Error Correction*Error Type*Subject	18	2.142		
Feedback*Error Correction*Error Type	4	0.260	0.27	0.895
Feedback*Error Correction*Error Type*Subject	32	0.964		

Table H-10. Analysis of variance table for in-vehicle task-function usability dependent variable, "Task Completion Time," and independent variables ASR system feedback modality, error correction method, and recognition error type.

Independent Variable	df	MS	F	p
Feedback	2	299.068	0.32	0.732
Feedback*Subject	18	942.103		
Error Correction	1	182.710	0.76	0.407
Error Correction*Subject	9	241.154		
Error Type	2	1581.647	20.75	0.0001
Error Type*Subject	18	76.218		
Feedback*Error Correction	2	209.092	0.65	0.537
Feedback*Error Correction*Subject	18	323.712		
Feedback*Error Type	4	108.438	0.85	0.501
Feedback*Error Type*Subject	36	126.966		
Error Correction*Error Type	2	619.255	8.21	0.003
Error Correction*Error Type*Subject	18	75.418		
Feedback*Error Correction*Error Type	4	57.018	0.53	0.716
Feedback*Error Correction*Error Type*Subject	34	108.084		

Table H-11. Analysis of variance table for driver preference/acceptance dependent variable, "Ease-of-Use," and independent variables ASR system feedback modality, error correction method, and recognition error type.

Independent Variable	df	MS	F	p
Feedback	2	1.410	0.68	0.517
Feedback*Subject	18	2.062		
Error Correction	1	1.747	1.96	0.195
Error Correction*Subject	9	0.890		
Error Type	2	62.940	20.62	0.001
Error Type*Subject	18	3.052		
Feedback*Error Correction	2	0.844	0.52	0.604
Feedback*Error Correction*Subject	18	1.630		
Feedback*Error Type	4	0.646	1.64	0.186
Feedback*Error Type*Subject	36	0.394		
Error Correction*Error Type	2	2.276	3.91	0.039
Error Correction*Error Type*Subject	18	0.583		
Feedback*Error Correction*Error Type	4	0.132	0.28	0.890
Feedback*Error Correction*Error Type*Subject	36	0.472		

Table H-12. Analysis of variance table for driver preference/acceptance dependent variable, “Comfort,” and independent variables ASR system feedback modality, error correction method, and recognition error type.

Independent Variable	df	MS	F	p
Feedback	2	1.527	1.17	0.333
Feedback*Subject	18	1.305		
Error Correction	1	0.747	0.54	0.481
Error Correction*Subject	9	1.381		
Error Type	2	12.299	13.75	0.0002
Error Type*Subject	18	0.895		
Feedback*Error Correction	2	0.414	0.31	0.734
Feedback*Error Correction*Subject	18	1.317		
Feedback*Error Type	4	0.213	0.52	0.719
Feedback*Error Type*Subject	36	0.406		
Error Correction*Error Type	2	1.006	1.41	0.270
Error Correction*Error Type*Subject	18	0.715		
Feedback*Error Correction*Error Type	4	0.872	2.30	0.080
Feedback*Error Correction*Error Type*Subject	36	0.380		

Table H-13. Analysis of variance table for driver preference/acceptance dependent variable, "Distraction," and independent variables ASR system feedback modality, error correction method, and recognition error type.

Independent Variable	df	MS	F	p
Feedback	2	6.066	1.97	0.168
Feedback*Subject	18	3.077		
Error Correction	1	2.216	1.34	0.277
Error Correction*Subject	9	1.652		
Error Type	2	70.300	43.88	0.0001
Error Type*Subject	18	1.602		
Feedback*Error Correction	2	5.674	3.08	0.072
Feedback*Error Correction*Subject	18	1.845		
Feedback*Error Type	4	1.726	2.13	0.097
Feedback*Error Type*Subject	36	0.810		
Error Correction*Error Type	2	2.334	2.47	0.113
Error Correction*Error Type*Subject	18	0.947		
Feedback*Error Correction*Error Type	4	0.198	0.45	0.773
Feedback*Error Correction*Error Type*Subject	36	0.441		

APPENDIX I: Analysis of Variance Tables for Experiment 2 Results

Table I-1. Analysis of variance table for driving performance dependent variable, "Mean Percent Eyes-Off-Road" and independent variables age, input condition, and task type.

Independent Variable	**df**	**MS**	**F**	**p**
Age	1	0.176	0.84	0.381
Subject(Age)	10	0.210		
Input Condition	7	1.240	16.34	0.0001
Input Condition*Subject(Age)	56	0.076		
Age*Input Condition	7	0.118	1.55	0.168
Input Condition*Subject(Age)	56	0.076		
Task*Input Condition	120	0.072	1.35	0.009
Task*Input Condition* Subject(Age)	1068	0.053		

Table I-2. Analysis of variance table for driving performance dependent variable, "Mean Percent Eyes-Off-Road" and independent variables age, recognition accuracy, and recognition error type.

Independent Variable	**df**	**MS**	**F**	**p**
Age	1	0.135	0.72	0.417
Subject(Age)	10	0.188		
Accuracy	2	0.164	6.28	0.008
Accuracy*Subject(Age)	20	0.026		
Error Type	1	0.680	4.79	0.065
Error Type*Subject(Age)	7	0.142		
Age*Accuracy	2	0.049	1.87	0.180
Accuracy*Subject(Age)	20	0.026		
Age*Error Type	1	0.181	1.28	0.296
Error Type*Subject(Age)	7	0.142		
Accuracy*Error Type	2	0.042	0.49	0.628
Accuracy*Error Type*Subject(Age)	11	0.086		
Age*Accuracy*Error Type	2	0.014	0.16	0.852
Accuracy*Error Type*Subject(Age)	11	0.086		

Table I-3. Analysis of variance table for driving performance dependent variable, "Steering Variability/Variance" and independent variables age, input condition, and task type.

Independent Variable	df	MS	F	p
Age	1	0.00614	53.35	0.0001
Subject(Age)	10	0.00012		
Input Condition	7	0.00008	0.85	0.550
Input Condition*Subject(Age)	70	0.00010		
Age*Input Condition	7	0.00011	1.10	0.376
Input Condition*Subject(Age)	70	0.00010		
Task*Input Condition	120	0.00009	0.92	0.708
Task*Input Condition* Subject(Age)	1292	0.00009		

Table I-4. Analysis of variance table for driving performance dependent variable, "Steering Variability/Variance," and independent variables age, recognition accuracy, and recognition error type.

Independent Variable	df	MS	F	p
Age	1	0.0045	22.28	0.0008
Subject(Age)	10	0.0002		
Accuracy	2	0.00016	2.00	0.162
Accuracy*Subject(Age)	20	0.000080		
Error Type	1	0.0000063	0.06	0.806
Error Type*Subject(Age)	10	0.000099		
Age*Accuracy	2	0.000048	0.60	0.561
Accuracy*Subject(Age)	20	0.000080		
Age*Error Type	1	0.000165	1.66	0.226
Error Type*Subject(Age)	10	0.000099		
Accuracy*Error Type	2	0.0000052	0.09	0.915
Accuracy*Error Type*Subject(Age)	20	0.000058		
Age*Accuracy*Error Type	2	0.000116	2.01	0.161
Accuracy*Error Type*Subject(Age)	20	0.000058		

Table I-5. Analysis of variance table for driving performance dependent variable, "Steering Reversals > 6°" and independent variables age, input condition, and task type.

Independent Variable	df	MS	F	p
Age	1	502.183	5.85	0.036
Subject(Age)	10	85.794		
Input Condition	7	66.146	3.82	0.001
Input Condition*Subject(Age)	70	17.294		
Age*Input Condition	7	26.086	1.51	0.179
Input Condition*Subject(Age)	70	17.294		
Task*Input Condition	120	26.379	2.72	0.0001
Task*Input Condition* Subject(Age)	1292	9.708		

Table I-6. Analysis of variance table for driving performance dependent variable, “Steering Reversals > 6°,” and independent variables age, recognition accuracy, and recognition error type.

Independent Variable	df	MS	F	p
Age	1	556.872	7.25	0.023
Subject(Age)	10	76.809		
Accuracy	2	160.220	13.24	0.0002
Accuracy*Subject(Age)	20	12.100		
Error Type	1	45.932	1.59	0.235
Error Type*Subject(Age)	10	28.806		
Age*Accuracy	2	28.126	2.32	0.124
Accuracy*Subject(Age)	20	12.100		
Age*Error Type	1	20.502	0.71	0.419
Error Type*Subject(Age)	10	28.806		
Accuracy*Error Type	2	13.791	1.21	0.320
Accuracy*Error Type*Subject(Age)	20	11.432		
Age*Accuracy*Error Type	2	0.624	0.05	0.947
Accuracy*Error Type*Subject(Age)	20	11.432		

Table I-7. Analysis of variance table for driving performance dependent variable, “Steering Reversals > 12°” and independent variables age, input condition, and task type.

Independent Variable	df	MS	F	p
Age	1	48.072	9.28	0.012
Subject(Age)	10	5.182		
Input Condition	7	5.192	2.92	0.010
Input Condition*Subject(Age)	70	1.780		
Age*Input Condition	7	2.242	1.26	0.283
Input Condition*Subject(Age)	70	1.780		
Task*Input Condition	120	2.529	1.53	0.0003
Task*Input Condition* Subject(Age)	1292	1.650		

Table I-8. Analysis of variance table for driving performance dependent variable, "Steering Reversals > 12°," and independent variables age, recognition accuracy, and recognition error type.

Independent Variable	df	MS	F	p
Age	1	52.955	7.72	0.020
Subject(Age)	10	6.862		
Accuracy	2	13.427	11.09	0.0006
Accuracy*Subject(Age)	20	1.211		
Error Type	1	3.341	2.31	0.159
Error Type*Subject(Age)	10	1.446		
Age*Accuracy	2	1.795	1.48	0.251
Accuracy*Subject(Age)	20	1.211		
Age*Error Type	1	5.388	3.73	0.082
Error Type*Subject(Age)	10	1.446		
Accuracy*Error Type	2	2.955	2.83	0.082
Accuracy*Error Type*Subject(Age)	20	1.043		
Age*Accuracy*Error Type	2	0.624	0.60	0.947
Accuracy*Error Type*Subject(Age)	20	1.043		

Table I-9. Analysis of variance table for driving performance dependent variable, “Accelerator Position Variability” and independent variables age, input condition, and task type.

Independent Variable	df	MS	F	p
Age	1	0.00624	37.73	0.0001
Subject(Age)	10	0.00016		
Input Condition	7	0.00035	1.83	0.095
Input Condition*Subject(Age)	70	0.00019		
Age*Input Condition	7	0.00035	1.81	0.099
Input Condition*Subject(Age)	70	0.00019		
Task*Input Condition	120	0.00013	1.02	0.419
Task*Input Condition* Subject(Age)	1292	0.00013		

Table I-10. Analysis of variance table for driving performance dependent variable, “Accelerator Position Variability,” and independent variables age, recognition accuracy, and recognition error type.

Independent Variable	df	MS	F	p
Age	1	0.00685	34.46	0.0002
Subject(Age)	10	0.00020		
Accuracy	2	0.00022	1.18	0.327
Accuracy*Subject(Age)	20	0.00019		
Error Type	1	0.000085	0.53	0.484
Error Type*Subject(Age)	10	0.000161		
Age*Accuracy	2	0.00026	1.36	0.280
Accuracy*Subject(Age)	20	0.00019		
Age*Error Type	1	0.000083	0.52	0.488
Error Type*Subject(Age)	10	0.000161		
Accuracy*Error Type	2	0.00029	1.25	0.307
Accuracy*Error Type*Subject(Age)	20	0.00023		
Age*Accuracy*Error Type	2	0.00017	0.75	0.486
Accuracy*Error Type*Subject(Age)	20	0.00023		

Table I-11. Analysis of variance table for driving performance dependent variable, “Mean Velocity” and independent variables age, input condition, and task type.

Independent Variable	df	MS	F	p
Age	1	0.000028	0.01	0.956
Subject(Age)	10	0.008798		
Input Condition	7	0.004533	1.15	0.342
Input Condition*Subject(Age)	70	0.003935		
Age*Input Condition	7	0.001442	0.37	0.919
Input Condition*Subject(Age)	70	0.003935		
Task*Input Condition	120	0.002800	0.79	0.948
Task*Input Condition* Subject(Age)	1292	0.003531		

Table I-12. Analysis of variance table for driving performance dependent variable, "Mean Velocity," and independent variables age, recognition accuracy, and recognition error type.

Independent Variable	df	MS	F	p
Age	1	0.000035	0.004	0.952
Subject(Age)	10	0.009185		
Accuracy	2	0.00375	1.27	0.303
Accuracy*Subject(Age)	20	0.00296		
Error Type	1	0.01653	1.70	0.222
Error Type*Subject(Age)	10	0.00975		
Age*Accuracy	2	0.00318	1.08	0.359
Accuracy*Subject(Age)	20	0.00296		
Age*Error Type	1	0.00042	0.04	0.840
Error Type*Subject(Age)	10	0.00975		
Accuracy*Error Type	2	0.00245	1.05	0.369
Accuracy*Error Type*Subject(Age)	20	0.00234		
Age*Accuracy*Error Type	2	0.00111	0.48	0.628
Accuracy*Error Type*Subject(Age)	20	0.00234		

Table I-13. Analysis of variance table for driving performance dependent variable, "Velocity Variability/Variance" and independent variables age, input condition, and task type.

Independent Variable	**df**	**MS**	**F**	**p**
Age	1	189054.847	1.47	0.253
Subject(Age)	10	128359.280		
Input Condition	7	61936.112	0.94	0.485
Input Condition*Subject(Age)	70	66167.990		
Age*Input Condition	7	66415.043	1.00	0.436
Input Condition*Subject(Age)	70	66167.990		
Task*Input Condition	120	1096.369	1.01	0.462
Task*Input Condition* Subject(Age)	1292	1087.880		

Table I-14. Analysis of variance table for driving performance dependent variable, "Velocity Variability/Variance," and independent variables age, recognition accuracy, and recognition error type.

Independent Variable	df	MS	F	p
Age	1	1671.076	0.93	0.357
Subject(Age)	10	1793.374		
Accuracy	2	705.678	1.38	0.275
Accuracy*Subject(Age)	20	512.542		
Error Type	1	267.961	0.62	0.449
Error Type*Subject(Age)	10	431.266		
Age*Accuracy	2	224.113	0.44	0.652
Accuracy*Subject(Age)	20	512.542		
Age*Error Type	1	768.375	1.78	0.212
Error Type*Subject(Age)	10	431.266		
Accuracy*Error Type	2	1351.043	2.04	0.157
Accuracy*Error Type*Subject(Age)	20	663.242		
Age*Accuracy*Error Type	2	73.385	0.11	0.896
Accuracy*Error Type*Subject(Age)	20	663.242		

Table I-15. Analysis of variance table for driving performance dependent variable, "Brake Activations" and independent variables age, input condition, and task type.

Independent Variable	df	MS	F	p
Age	1	665.719	35.74	0.0001
Subject(Age)	10	18.627		
Input Condition	7	8.883	1.50	0.182
Input Condition*Subject(Age)	70	5.923		
Age*Input Condition	7	13.482	2.28	0.038
Input Condition*Subject(Age)	70	5.923		
Task*Input Condition	120	3.373	3.07	0.0001
Task*Input Condition* Subject(Age)	1292	1.099		

Table I-16. Analysis of variance table for driving performance dependent variable, "Brake Activations," and independent variables age, recognition accuracy, and recognition error type.

Independent Variable	df	MS	F	p
Age	1	709.709	66.09	0.0001
Subject(Age)	10	10.738		
Accuracy	2	1.448	0.73	0.496
Accuracy*Subject(Age)	20	1.997		
Error Type	1	23.098	1.55	0.242
Error Type*Subject(Age)	10	14.923		
Age*Accuracy	2	4.527	2.27	0.130
Accuracy*Subject(Age)	20	1.997		
Age*Error Type	1	7.823	0.52	0.486
Error Type*Subject(Age)	10	14.923		
Accuracy*Error Type	2	4.172	2.71	0.091
Accuracy*Error Type*Subject(Age)	20	1.538		
Age*Accuracy*Error Type	2	1.492	0.97	0.396
Accuracy*Error Type*Subject(Age)	20	1.538		

Table I-17. Analysis of variance table for in-vehicle task-function usability dependent variable, "Task Completion Time," and independent variables age, input condition, and task type.

Independent Variable	df	MS	F	p
Age	1	359.827	3.35	0.097
Subject(Age)	10	107.438		
Input Condition	7	882.850	8.35	0.0001
Input Condition*Subject(Age)	70	105.730		
Age*Input Condition	7	48.716	0.46	0.860
Input Condition*Subject(Age)	70	105.730		
Task*Input Condition	120	520.303	11.28	0.0001
Task*Input Condition* Subject(Age)	1292	46.129		

Table I-18. Analysis of variance table for in-vehicle task-function usability dependent variable, "Task Completion Time," and independent variables age, recognition accuracy, and recognition error type.

Independent Variable	df	MS	F	p
Age	1	272.360	3.55	0.089
Subject(Age)	10	76.788		
Accuracy	2	836.135	6.52	0.007
Accuracy*Subject(Age)	20	128.340		
Error Type	1	623.031	4.37	0.063
Error Type*Subject(Age)	10	142.420		
Age*Accuracy	2	10.246	0.08	0.924
Accuracy*Subject(Age)	20	128.340		
Age*Error Type	1	139.649	0.98	0.345
Error Type*Subject(Age)	10	142.420		
Accuracy*Error Type	2	420.460	6.68	0.006
Accuracy*Error Type*Subject(Age)	20	62.913		
Age*Accuracy*Error Type	2	92.326	1.47	0.254
Accuracy*Error Type*Subject(Age)	20	62.913		

Table I-19. Analysis of variance table for driver preference/acceptance dependent variable, "Ease-of-Use," and independent variables age, input condition, and task type.

Independent Variable	df	MS	F	p
Age	1	9.935	0.29	0.602
Subject(Age)	10	34.177		
Input Condition	7	4.922	4.19	0.0007
Input Condition*Subject(Age)	70	1.174		
Age*Input Condition	7	0.900	0.77	0.617
Input Condition*Subject(Age)	70	1.174		
Task*Input Condition	120	0.827	6.23	0.0001
Task*Input Condition* Subject(Age)	1292	0.133		

Table I-20. Analysis of variance table for driver preference/acceptance dependent variable, "Ease-of-Use," and independent variables age, recognition accuracy, and recognition error type.

Independent Variable	df	MS	F	p
Age	1	13.277	0.44	0.524
Subject(Age)	10	30.421		
Accuracy	2	12.046	9.89	0.001
Accuracy*Subject(Age)	20	1.218		
Error Type	1	0.783	0.44	0.524
Error Type*Subject(Age)	10	1.792		
Age*Accuracy	2	0.416	0.34	0.715
Accuracy*Subject(Age)	20	1.218		
Age*Error Type	1	0.120	0.07	0.801
Error Type*Subject(Age)	10	1.792		
Accuracy*Error Type	2	0.267	0.59	0.563
Accuracy*Error Type*Subject(Age)	20	0.451		
Age*Accuracy*Error Type	2	0.599	1.33	0.288
Accuracy*Error Type*Subject(Age)	20	0.451		

Table I-21. Analysis of variance table for driver preference/acceptance dependent variable, “Comfort,” and independent variables age, input condition, and task type.

Independent Variable	df	MS	F	p
Age	1	0.016	0.0005	0.982
Subject(Age)	10	30.826		
Input Condition	7	4.723	4.86	0.0002
Input Condition*Subject(Age)	70	0.972		
Age*Input Condition	7	0.787	0.81	0.582
Input Condition*Subject(Age)	70	0.972		
Task*Input Condition	120	0.904	6.88	0.0001
Task*Input Condition*Subject(Age)	1292	0.131		

Table I-22. Analysis of variance table for driver preference/acceptance dependent variable, "Comfort," and independent variables age, recognition accuracy, and recognition error type.

Independent Variable	df	MS	F	p
Age	1	0.045	0.002	0.969
Subject(Age)	10	27.414		
Accuracy	2	10.378	10.52	0.0008
Accuracy*Subject(Age)	20	0.986		
Error Type	1	3.430	2.62	0.136
Error Type*Subject(Age)	10	1.308		
Age*Accuracy	2	0.695	0.70	0.506
Accuracy*Subject(Age)	20	0.986		
Age*Error Type	1	0.073	0.06	0.818
Error Type*Subject(Age)	10	1.308		
Accuracy*Error Type	2	0.355	0.78	0.470
Accuracy*Error Type*Subject(Age)	20	0.453		
Age*Accuracy*Error Type	2	1.172	2.59	0.100
Accuracy*Error Type*Subject(Age)	20	0.453		

Table I-23. Analysis of variance table for driver preference/acceptance dependent variable, "Distraction," and independent variables age, input condition, and task type.

Independent Variable	df	MS	F	p
Age	1	10.017	0.50	0.497
Subject(Age)	10	20.148		
Input Condition	7	12.596	8.37	0.0001
Input Condition*Subject(Age)	70	1.506		
Age*Input Condition	7	0.683	0.45	0.864
Input Condition*Subject(Age)	70	1.506		
Task*Input Condition	120	2.119	12.86	0.0001
Task*Input Condition* Subject(Age)	1292	0.165		

Table I-24. Analysis of variance table for driver preference/acceptance dependent variable, "Distraction," and independent variables age, recognition accuracy, and recognition error type.

Independent Variable	df	MS	F	p
Age	1	6.985	0.34	0.573
Subject(Age)	10	20.554		
Accuracy	2	20.072	20.93	0.0001
Accuracy*Subject(Age)	20	0.959		
Error Type	1	17.753	14.34	0.004
Error Type*Subject(Age)	10	1.238		
Age*Accuracy	2	0.698	0.73	0.496
Accuracy*Subject(Age)	20	0.959		
Age*Error Type	1	3.251	2.63	0.136
Error Type*Subject(Age)	10	1.238		
Accuracy*Error Type	2	1.442	2.44	0.112
Accuracy*Error Type*Subject(Age)	20	0.590		
Age*Accuracy*Error Type	2	0.014	0.02	0.976
Accuracy*Error Type*Subject(Age)	20	0.590		

VITA

Andrew William Gellatly

1871 St. Andrews Circle
Blacksburg, Virginia 24060
(540) 552-8832
email: gellatly@vt.edu

EDUCATION

Ph.D. in Industrial and Systems Engineering (Human Factors Option), May 1997
Virginia Polytechnic Institute and State University — Blacksburg, VA
GPA: 3.91/4.00

M.S. in Industrial and Systems Engineering (Human Factors & Safety Option), May 1995
Virginia Polytechnic Institute and State University — Blacksburg, VA

Post-Graduate Studies in Mathematics, August 1990 to May 1991
Washtenaw Community College — Ann Arbor, MI

B.S. degree in Psychology, December 1989
The University of Michigan — Ann Arbor, MI

WORK EXPERIENCE

Graduate Research Assistant — November 1995 - February 1997
Virginia Polytechnic Institute and State University — Blacksburg, VA

Performed research in the areas of speech recognition technology and automotive secondary-control systems. Developed decision tools for the allocation of in-vehicle tasks to either speech control or manual control by the driver of a vehicle. Designed experiments to evaluate the effects of various automatic speech recognition system parameters (recognition errors, feedback, error correction methods, and recognition accuracy), input modalities (manual or speech), and age on driving performance, system usability, and user preference/acceptance. Developed models, recommendations, and guidelines to be used by designers and engineers of automotive secondary-control systems from the results of the empirical research.

User Interface Development Summer Intern — May 1995 - October 1995
Delco Electronics Corporation — Kokomo, IN

Provided technical support for a project aimed at examining potential driver performance benefits of advanced in-vehicle displays and controls. Primarily responsible for executing the experimental design, experimental testing, and data collection. Also responsible for instrumenting two automobiles for data collection and programming PC-based data collection software. Concluded with conducting in-traffic data collection. Supported product designers and engineers with human factors recommendations and guidelines for system development including display design and control labeling.

Graduate Teaching Assistant | August 1991 - May 1995
Virginia Polytechnic Institute and State University | Blacksburg, VA

Assisted in the instruction of undergraduate industrial engineering classes covering occupational safety and hazard control and industrial fire control, and a graduate level industrial engineering class on system safety analysis. Attended lectures and recorded detailed notes. Prepared, administered, and graded exams. Instructed class in the absence of the professor. Held office hours and answered student questions and concerns.

User Interface Development Summer Intern | May 1994 - August 1994
Delco Electronics Corporation | Kokomo, IN

Provided technical support to a senior researcher for a project aimed at examining potential driver performance benefits of head-up displays. Conducted both a closed-course field study and an in-traffic study. Assisted in the execution of the experimental design, experimental testing, and data collection. Analyzed preliminary findings from the data using PC-based statistical software.

Human Factors Research Associate I | May 1992 - August 1992
The University of Michigan Transportation Research Institute | Ann Arbor, MI

Performed research on effects of foreground illumination in automobile headlighting and driver eye-movements using rear-vision mirror systems. Executed experimental design, set-up, and data collection. Conducted statistical data analysis and assisted in writing technical reports. Also assisted in the publication of journal articles by colleagues in the department.

Human Factors Research Assistant II | February 1990 - August 1991
The University of Michigan Transportation Research Institute | Ann Arbor, MI

Aided in the design and development of research experiments to improve transportation safety. Conducted research on vehicle lighting, vehicle conspicuity, glare scaling techniques, traffic signing, and rear vision systems. Collected data from experiments and performed statistical analyses using computer software. Published technical reports and journal articles from the research conducted. Performed independent research on topics related to improving existing experimental procedures.

Officer Candidate United States Marine Corps | November 1987 - May 1990
Marine Corps Development and Education Command | Quantico, VA

Completed training for Platoon Leaders Class (Junior). Held leadership billets placing me in charge of as few as four candidates and up to as many as sixty candidates. Awarded Superior Physical Performance Certificate. Elected to disenroll due to knee surgery.

ACADEMIC AWARDS/HONORS

ALPHA PI MU Membership September 1992 - present
Membership in the only nationally recognized industrial engineering honor society awarded for academic achievement.

James Langston Parker Memorial Scholarship December 1992 and September 1996
Scholarship awarded by the College of Engineering at Virginia Polytechnic Institute and State University in recognition of academic performance and leadership ability in the field of safety engineering.

National Safety Council Public Utilities Section Scholarship July 1994
Scholarship awarded by the Public Utilities Section of the National Safety Council in recognition of academic performance and interest in the safety and health profession.

PROFESSIONAL MEMBERSHIPS

Aerospace Medical Association (AsMA) Member October 1993 - October 1995
American Society of Safety Engineers (ASSE) Member November 1991 - May 1995
Human Factors and Ergonomics Society (HFES) Member July 1990 - present
- •Aerospace Technical Group Member
- •Consumer Products Technical Group Member
- •Safety Technical Group Member
- •Surface Transportation Technical Group Member
- •Visual Performance Technical Group Member

System Safety Society Member October 1993 - October 1995
Member of Virginia Tech Student Chapter of HFS September 1991 - present
Vice-president of Virginia Tech Student Chapter of HFS April 1992 - December 1992
Member of Virginia Tech Student Chapter of ASSE November 1991 - May 1996
Michigan Licensed Builder April 1990 - May 1992

PUBLICATIONS

Dingus, T.A., Gellatly, A.W., and Reinach, S. (in press). Human-computer interaction applications for intelligent transportation systems. In M. Helander (Ed.), Handbook of Human-Computer Interaction (2nd. Ed.).

Kiefer, R.J., and Gellatly, A.W. (1996). Toward quantifying the real-world consequences of the claimed "eyes-on-road" benefit of head-up displays (SAE Technical Paper Series No. 960946). Warrendale, PA: Society of Automotive Engineers.

Flannagan, M.J., Sivak, M., and Gellatly, A.W. (1992). Rearview mirror reflectivity and the tradeoff between forward and rearward vision (SAE Technical Paper Series No. 920404). Warrendale, PA: Society of Automotive Engineers.

Sivak, M., Flannagan, M., Chandra, D., and Gellatly, A.W. (1992). Visual aiming of European and U.S. low-beam headlamps (SAE Technical Paper Series No. 920814). Warrendale, PA: Society of Automotive Engineers.

Sivak, M., Flannagan, M., and Gellatly, A.W. (1992). The influence of truck driver eye position on the effectiveness of retroreflective traffic signs. Lighting Research and Technology, 25, 31-36.

Sivak, M., Flannagan, M., Gellatly, A.W., and Luoma, J. (1992). Subjective preferences for the red color of stop signs: Implications for the design of high-intensity-discharge headlamps. Color Research and Application, 17, 356-360.

Flannagan, M.J., Sivak, M., and Gellatly, A.W. (1991). Rearward vision, driver confidence, and discomfort glare using an electrochromic rearview mirror (SAE Technical Paper Series No. 910822). In D. Hoffmeister (Ed.), Vehicle Lighting Design for Optimal Visibility and Performance: SP-857 (pp. 25-40). Warrendale, PA: Society of Automotive Engineers.

Sivak, M., Flannagan, M., and Gellatly, A.W. (1991). Reaction times to high-contrast brake lamps (SAE Technical Paper Series No. 910821). In D. Hoffmeister (Ed.), Vehicle Lighting Design for Optimal Visibility and Performance: SP-857 (pp. 19-23). Warrendale, PA: Society of Automotive Engineers.

TECHNICAL REPORTS

Gellatly, A.W. (1995). Effects of seated posture on static strength, lower-body isometric muscle contractions, and manual tracking performance. Unpublished masters thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Gellatly, A.W., and Weintraub, D.J. (1990). User reconfigurations of the de Boer rating scale for discomfort glare. Michigan University, Ann Arbor, Transportation Research Institute. 21 p. Sponsor: Michigan University, Ann Arbor, Industry Affiliation Program for Human Factors in Transportation Safety. Report No. UMTRI-90-20.

Sivak, M., Gellatly, A.W., and Flannagan, M. (1991). Minimum light above horizontal of low-beam headlamps for nighttime legibility of traffic signs. Michigan University, Ann Arbor, Transportation Research Institute. 33 p. Sponsor: Motor Vehicle Manufacturers Association, Detroit, Michigan. Report No. UMTRI-91-3.

Weintraub, D.J., Gellatly, A.W., Sivak, M., and Flannagan, M. (1991). Methods for evaluating discomfort glare. Michigan University, Ann Arbor, Transportation Research Institute. 21 p. Sponsor: Michigan University, Ann Arbor, Industry Affiliation Program for Human Factors in Transportation Safety. Report No. UMTRI-91-13.

Flannagan, M.J., Sivak, M., Gellatly, A.W., and Luoma, J. (1992). A field study of discomfort glare from high-intensity discharge headlamps. Michigan University, Ann Arbor, Transportation Research Institute. 12 p. Sponsor: Michigan University, Ann Arbor, Industry Affiliation Program for Human Factors in Transportation Safety. Report No. UMTRI-92-16.

Flannagan, M.J., Sivak, M., Gellatly, A.W., and Luoma, J. (1992). Ranges of stop sign chromaticity under tungsten-halogen and high-intensity discharge illumination. Michigan University, Ann Arbor, Transportation Research Institute. 19 p. Sponsor: Michigan University, Ann Arbor, Industry Affiliation Program for Human Factors in Transportation Safety. Report No. UMTRI-92-17.

Flannagan, M.J., Sivak, M., and Gellatly, A.W. (1991). Joint effects of wavelength and ambient luminance on discomfort glare from monochromatic and bichromatic sources. Michigan University, Ann Arbor, Transportation Research Institute. 39 p. Sponsor: Michigan University, Ann Arbor, Industry Affiliation Program for Human Factors in Transportation Safety. Report No. UMTRI-91-42.

Flannagan, M.J., Sivak, M., and Gellatly, A.W. (1991). Rearview mirror reflectivity and the tradeoff between forward and rearward vision. Michigan University, Ann Arbor, Transportation Research Institute. 23 p. Sponsor: Ichikoh Industries Ltd., Tokyo, Japan. Report No. UMTRI-91-47.

Sivak, M., Flannagan, M., Chandra, D., and Gellatly, A.W. (1991). Visual aiming of European and U.S. low-beam headlamps. Michigan University, Ann Arbor, Transportation Research Institute. 30 p. Sponsor: Motor Vehicle Manufacturers Association, Detroit, Michigan. Report No. UMTRI-91-34.

Sivak, M., Flannagan, M., and Gellatly, A.W. (1991). The influence of truck driver eye position on the effectiveness of retroreflective traffic signs. Michigan University, Ann Arbor, Transportation Research Institute. 19 p. Sponsor: Michigan University, Ann Arbor, Industry Affiliation Program for Human Factors in Transportation Safety. Report No. UMTRI-91-35.

Sivak, M., Flannagan, M., Gellatly, A.W., and Luoma, J. (1991). Subjective preferences for the red color of stop signs: Implications for the design of high-intensity-discharge headlamps. Michigan University, Ann Arbor, Transportation Research Institute. 19 p. Sponsor: Michigan University, Ann Arbor, Industry Affiliation Program for Human Factors in Transportation Safety. Report No. UMTRI-91-27.

Flannagan, M.J., Sivak, M., and Gellatly, A.W. (1990). Rearward vision, driver confidence, and discomfort glare using an electrochromic rearview mirror. Michigan University, Ann Arbor, Transportation Research Institute. 29 p. Sponsor: Ichikoh Industries Ltd., Tokyo, Japan. Report No. UMTRI-90-27.

Sivak, M., Flannagan, M., and Gellatly, A.W. (1990). Reaction times to high-contrast brake lamps. Michigan University, Ann Arbor, Transportation Research Institute. 11 p. Sponsor: Michigan University, Ann Arbor, Industry Affiliation Program for Human Factors in Transportation Safety. Report No. UMTRI-90-30.

Andrew William Gellatly